A User-Guided Generation Framework for Personalized Music Synthesis Using Interactive Evolutionary Computation

Yanan Wang
University of Aizu
Aizuwakamatsu, Fukushima, Japan
d8232116@u-aizu.ac.jp

Yan Pei*
University of Aizu
Aizu-Wakamatsu, Fukushima, Japan
peiyan@u-aizu.ac.jp

Zerui Ma
Beijing University of Technology
Beijing, China
mazerui@emails.bjut.edu.cn

Jianqiang Li
Beijing University of Technology
Beijing, China
lijianqiang@bjut.edu.cn

ABSTRACT

The development of generative artificial intelligence (AI) has demonstrated notable advancements in the domain of music synthesis. However, a perceived lack of creativity in the generated content has drawn significant attention from the public. To address this, this paper introduces a novel approach to personalized music synthesis, incorporating a human-in-the-loop generation. This method leverages the dual strengths of interactive evolutionary computation, known for its capturing user preferences, and generative adversarial network, renowned for its capacity to autonomously produce high-quality music. The primary objective of this integration is to augment the credibility and diversity of generative AI in music synthesis, fostering computational artistic creativity in humans. Furthermore, a user-friendly interactive music player has been designed to facilitate users in the music synthesis process. The proposed method exemplifies a paradigm wherein users manipulate latent space through human-machine interaction, underscoring the pivotal role of humans in the synthesis of diverse and creative music.

CCS CONCEPTS

• **Theory of computation** → *Interactive computation*; • **Human-centered computing** → *Interaction design theory, concepts and paradigms*; • **Applied computing** → *Sound and music computing*; • **Computing methodologies** → *Neural networks*.

KEYWORDS

Human-AI interaction, Co-creativity, Music synthesis, Interactive evolutionary computation, Generative adversarial network, Human-in-the-loop.

ACM Reference Format:
Yanan Wang, Yan Pei*, Zerui Ma, and Jianqiang Li. 2024. A User-Guided Generation Framework for Personalized Music Synthesis Using Interactive Evolutionary Computation. In *Genetic and Evolutionary Computation Conference (GECCO '24 Companion), July 14–18, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3638530.3664109

1 INTRODUCTION

Generative artificial intelligence (AI) stands as a transformative technology, which is capable of generating novel content by learning and mining existing data through a series of machine learning algorithms and deep learning models. Presently, it finds widespread application across diverse domains, including education [3], art [11], medicine [12], video [1], and image [32], etc. Notably, in the realm of music, various models such as autoregressive [16, 29, 48], generative adversarial network (GAN) [9, 19, 24], diffusion [2, 5, 13, 23], and other models [8] have become prevalent for music synthesis.

One of the primary challenges currently faced by generative music models is the perceived lack of creativity and personalization, as highlighted by several studies [7, 30]. This challenge is principally rooted in the opaque nature of end-to-end learning in generative AI models, limiting user engagement in the generation process. The black-box nature impedes the model's ability to promptly comprehend user preferences and creative inspirations, ultimately hindering its capacity to produce personalized and innovative music. Moreover, there is currently a lack of methods for human-computer interaction to manipulate the potential space within deep learning and guide variables to the areas of interest for users. To address these issues, there is a critical need to establish an interactive, human-centered, and co-creative generative model that empowers users to contribute to the generation process, resulting in diverse and personalized musical outputs. Fortunately, the expeditious realization of this objective has been made possible through the application of interactive evolutionary computation (IEC). This powerful methodology utilizes interactive technologies to explore users' subjective preferences, guiding evolutionary computation algorithms to optimize the target system.

In response to these challenges, we propose a user-guided generation framework for personalized music synthesis. This method capitalizes on the dual strengths of IEC, known for its proficiency in capturing user preferences, and GAN, celebrated for its autonomous

*Dr. Yan Pei is the corresponding author. https://www.u-aizu.ac.jp/~peiyan/.

GECCO '24 Companion, July 14–18, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0495-6/24/07…$15.00
https://doi.org/10.1145/3638530.3664109

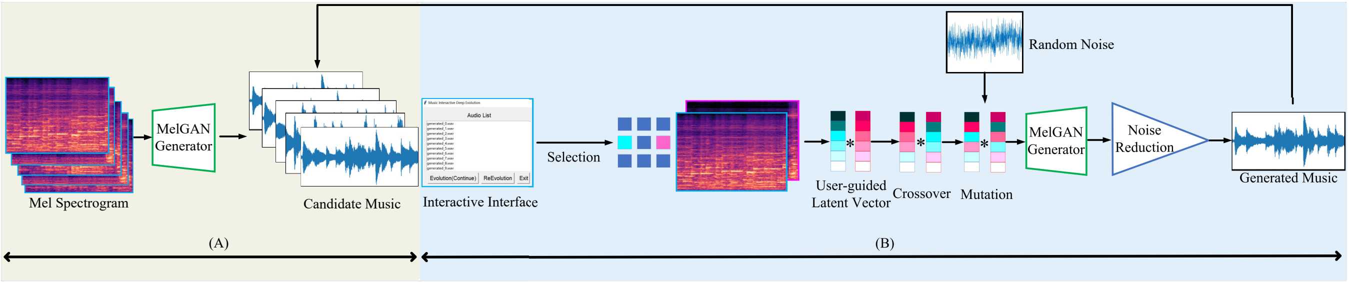

Figure 1: Overview of UIGAN. The framework comprises two primary components: (A) the MelGAN-based pre-trained generation, and (B) the interactive evolutionary manipulation.

generation of high-quality music. The primary objective of this integration is to augment the credibility and diversity of generative AI in music synthesis, thereby igniting computational artistic creativity in humans. Our proposed model underwent comprehensive comparative experiments utilizing a piano dataset, demonstrating its capability to generate diverse and creative solutions from both quantitative and qualitative perspectives. Moreover, a user-friendly interactive music player has been designed to facilitate users in the music synthesis process. This paradigm illustrates a transformative framework wherein users actively manipulate latent space through human-machine interaction, emphasizing the indispensable role of humans in synthesizing diverse and creative music.

The structure of the paper is organized as follows. In this section (Section 1), we provide an overview of the prevailing landscape of generative AI, elucidating the motivation behind our research. Section 2 delves into a comprehensive review of prior research in the realms of music synthesis and IEC. The intricacies of our proposed model and associated details are expounded upon in Section 3. Subsequently, Section 4 outlines the methodology employed in the comparative experiment conducted. Finally, the findings of the study are encapsulated in Section 5, which serves to summarize outcomes and proffer potential avenues for future research.

2 RELATED WORKS AND TECHNOLOGIES

2.1 Music Synthesis

Music synthesis constitutes a generative task wherein a model produces coherent music samples. This task is categorized into conditional and unconditional synthesis based on its dependence on conditional inputs. Four dominant categories in this field include autoregressive models such as WaveNet [48], SampleRNN [29], and WaveRNN [16]; GAN-based models like WaveGAN [9] and MelGAN [19]; VQ-VAE models exemplified by Jukebox [8]; and diffusion models, including Diffsinger [23].

In particular, GAN-based models [9, 19] leverage concealed latent spaces to map audio feature sequences, enabling the exploration of music's latent features and enhancing its diversity. Additionally, modeling mel-spectrograms not only streamlines the overarching temporal and spectral structure but also constitutes conditional generation representing the user's initial synthesis preferences [19, 24]. Therefore, MelGAN emerges as a suitable candidate for interactive music synthesis.

The manipulation of latent space in audio generative models presents a promising area of research. While unsupervised methods like dimensionality reduction [43] and smoothing [51] have been explored, their limitation lies in the inability to precisely focus latent vectors on the user's region of interest. To address this gap, an urgently needed interactive supervised method is discussed in this study, employing a human-machine interaction approach that successfully concentrates latent vectors toward the desired user-centric regions.

2.2 Interactive Evolutionary Computation

IEC has found wide-ranging applications across diverse domains, including jewelry design [50], pattern design [53], dance design [10, 36], fractal modeling [33], color palette design [28], fashion design [41], image generation [4, 52], image retrieval systems [6], texture modeling [26], hearing aid fitting [44], sound composition [49], and more.

However, in contrast to the progress observed in computer vision and art design, advancements in IEC for music waveform synthesis have been comparatively limited. Many efforts in music synthesis and composition rely on musical instrument digital interface (MIDI) synthesizers, necessitating professional music knowledge for manipulating various parameters [25, 47, 49]. Recognizing this gap, it is imperative to develop a music waveform synthesis method that seamlessly integrates IEC's capability to capture user preferences with generative AI's capacity for generating high-quality solutions. This integration is vital for enhancing the diversity of synthesized music, as emphasized in [34].

3 THE PROPOSED FRAMEWORK: UIGAN FOR PERSONALIZED MUSIC SYNTHESIS USING INTERACTIVE EVOLUTIONARY COMPUTATION

This section introduces the user-interactive generative adversarial network (UIGAN) for music synthesis, building upon the foundation of the existing music waveform synthesis model MelGAN [19]. UIGAN aims to evolve and manipulate the latent vector of MelGAN, capturing users' auditory preferences through human-machine interaction.

The model comprises two primary components, as illustrated in Figure. 1: (A) the MelGAN-based pre-trained generation, and (B) the interactive evolutionary manipulation. In the initial phase, the

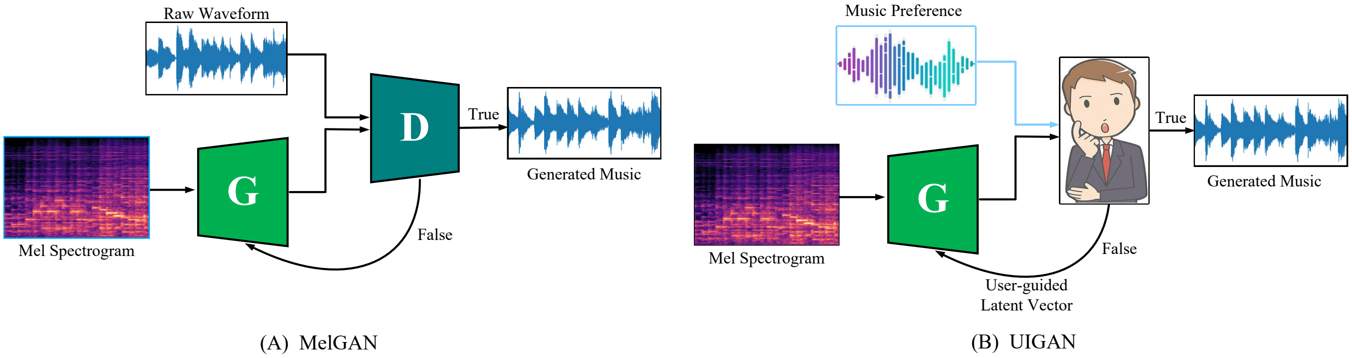

Figure 2: Overall structure of MelGAN and UIGAN. MelGAN is structured with a discriminator (D) and a generator (G). Similar to the discriminator's role in MelGAN, humans assume a crucial role in UIGAN, actively filtering out low-quality music and selecting musical pieces that resonate with their auditory preferences. The integration of human subjective evaluation within UIGAN represents an innovative approach to collaborate between deep learning-based generation models and human input, thereby facilitating the generation of personalized content with human-in-the-loop.

MelGAN vocoder undergoes pre-training to generate audio latent representations. The pre-trained MelGAN model produces candidate music individuals by mapping the mel-spectrograms of the music selected by the user for adaptation.

Users engage with an interactive interface to express their auditory preferences and choose their preferred music. The mel-spectrograms corresponding to the selected music serve as user-guided latent vectors for both subsequent evolutionary manipulations and MelGAN regeneration.

3.1 MelGAN Generation

We introduce the generator, discriminator and training objective of the MelGAN vocoder to enhance understanding of our proposed framework.

3.1.1 Generator. The generator is constructed using a fully convolutional feed-forward network, taking a mel-spectrogram as its input and producing an audio signal as its output. The overall generation network consists of two one-dimensional convolutional layers with a kernel size of 7, four up-sampling layers, and four residual blocks. This architecture employs transposed convolutional layers for up-sampling, followed by residual blocks incorporating dilated convolutions. Each residual layer is comprised of three residual blocks with dilation factors of 1, 3, and 9, respectively. All layers utilize a leaky ReLU activation function and weight normalization [39] to ensure optimal performance. During the training of the generator, feature matching [21] is applied to minimize the $L1$ distance, contributing to the overall enhancement of performance.

3.1.2 Discriminator. The discriminator in MelGAN adopts a multi-scale architecture, which comprises three discriminators with different frequency ranges and audio scales (denoted as D_1, D_2, and D_3). D_1 assesses the original music, while D_2 and D_3 operate on downsampled audio at scales 2 and 4, respectively. Additionally, a PatchGAN-like Markov window discriminator [15] is incorporated to construct discriminators at different frequency scales. Each self-discriminator consists of four downsampling layers with a scale of 4 and three one-dimensional convolutional layers using convolution kernels of 15, 5, and 3. Similar to the generator, each layer of the discriminator is equipped with weight normalization and a leaky ReLU activation function. For specific parameter settings and training details, please refer to the original paper on MelGAN by Kumar et al. [19].

3.1.3 Training Objective. Hinge loss function [22] was applied to train the discriminator.

$$\min_{D_k} \mathbb{E}_x \left[\min\left(0, 1 - D_k(x)\right)\right] + \mathbb{E}_{s,z}\left[\min\left(0, 1 + D_k(G(s,z))\right)\right], \quad \forall k = 1,2,3 \qquad (1)$$

$$\min_G \mathbb{E}_{s,z}\left[\sum_{k=1,2,3} -D_k(G(s,z))\right] \qquad (2)$$

where D_k is the k^{th} discriminator, and x, s, z represent the raw waveform, input mel-spectrogram, and Gaussian noise respectively.

Feature match [21] was used to train the generator for minimizing $L1$ distance. This distance represents differences between the discriminator feature maps of real and synthetic audio.

$$\mathcal{L}(G, D_k) = \mathbb{E}_{x,s}\left[\sum_{i=1}^{T} \frac{1}{N_i} \left\| D_k^{(i)}(x) - D_k^{(i)}(G(s,z)) \right\|_1\right] \qquad (3)$$

$D_k^{(i)}$ represents the feature map output of the i^{th} layer from the k^{th} discriminator block, and N_i is the number of units in each layer.

The final objective of MelGAN with $\lambda = 10$ is:

$$\min_G \left(\mathbb{E}_{s,z}\left[\sum_{k=1}^{K} \left(D_k(G(s,z)) - 1\right)^2\right] + \lambda \sum_{k=1}^{K} \mathcal{L}(G, D_k)\right) \qquad (4)$$

Algorithm 1 shows the training process of MelGAN.

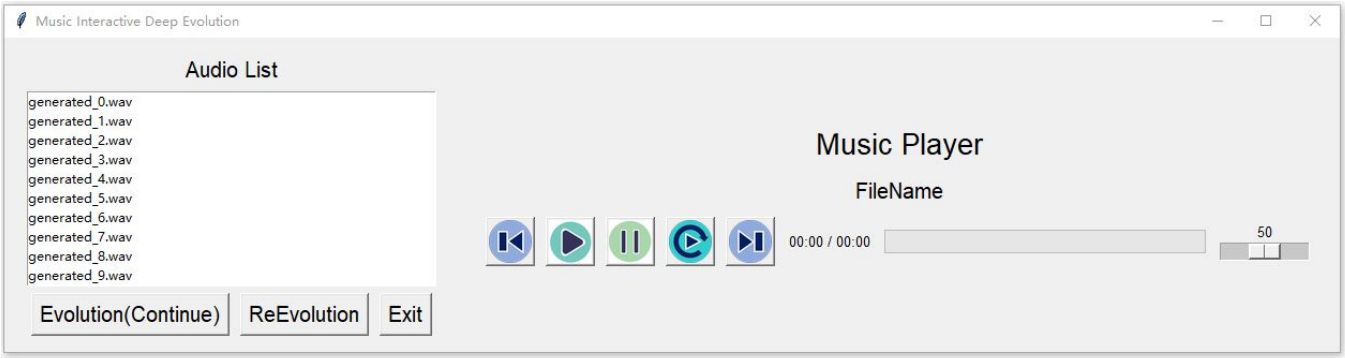

Figure 3: Interface of the designed interactive music player. The functionalities of the music player are organized into two main categories: interactive evolutionary manipulation and musical playback. Interactive evolutionary manipulation comprises evolution (continue), re-evolution, and exit. This function involves actively shaping and refining the music selection process. On the other hand, the music playback function provides options for playing the previous track, playing, pausing, replaying, and advancing to the next track. This component focuses on the seamless enjoyment and control of the music playback experience.

Algorithm 1 Training MelGAN model

```
Require: Training audio dataset $X$, Number of iterations $T$, Convergence criteria threshold $\epsilon$, Generator network $G$, Discriminator network $D(k), k = 1, 2, 3$.
Ensure: Trained generator $G$ and discriminator $D$.
Initialize the parameters of MelGAN
for $t = 1$ to $T$ do
    Sample a mini-batch of real audio samples $\{x_1, x_2, ..., x_N\}$ from $X$.
    Calculate the discriminator loss $D_k$ using the Hinge loss function in Eq. (1).
    Update the discriminator parameters $D_k$.
    Calculate the generator loss using feature matching in Eq. (4).
    Update the generator parameters $G$.
    if MelGAN loss change $< \epsilon$ or $t == T$ then
        break
    end if
end for
```

3.2 Interactive Evolutionary Manipulation

Interactive evolutionary manipulation within UIGAN involves a sequence of operations executed through IEC. This process guides user-driven latent vectors within the search space toward the user's specific region of interest. These operations include various IEC components and a regeneration component. The former consists of interactive selection, crossover, and mutation. Interactive selection plays a pivotal role in capturing user preferences, while crossover and mutation serve as methods to further integrate and enhance the user-guided latent vectors. After these optimization operations, regeneration is employed to further synthesize music. This process utilizes personalized, user-guided vectors or mel-spectrograms as inputs to a pre-trained MelGAN model. The candidate music undergoes continuous regeneration through interactive evolutionary manipulation operations until it generates a piece of music that resonates with the user's preferences.

3.2.1 Interactive Selection Operation. The interactive selection operation determines evolved parents and generates user-guided latent vectors based on user auditory preferences. Users only need to select their favorite musical pieces through an interactive music player (Figure. 3) to perform this operation. The latent vectors corresponding to the selected pieces through interactive selection can accurately encapsulate the user's auditory preferences. Interactive selection and fitness approximation are currently common methods used in IEC to capture user preferences [49]. Fitness approximation involves predicting or estimating individual fitness values to guide the progress of the evolutionary algorithm [37, 41]. In comparison to fitness approximation, interactive selection exhibits pronounced personalization and targeting [4, 27, 45, 49]. This operation is similar to the mechanism of paired comparison in interactive differential evolution and can effectively reduce user fatigue and obtain user auditory preferences [35, 45].

Moreover, users play a pivotal role in this selection process, undertaking responsibilities such as retaining high-quality music and filtering out music they dislike. They not only successfully guide latent vectors toward their areas of interest but also ensure the quality of vocoder music synthesis. The role of humans in UIGAN resembles that of discriminators in GAN, with tangible effects being more intelligent content generation and the production of personalized content (Figure. 2).

3.2.2 Crossover Operation. Crossover is a recombination process designed to amalgamate selected parent individuals or musical pieces, thereby generating new individual vectors. This process holds a pivotal role in preserving population diversity and fostering evolution through the amalgamation of bi-parental vectors. The uniform crossover method [42] is utilized in this study, acknowledged for adeptly merging attributes from both parents with a fixed probability.

3.2.3 Mutation Operation. Mutation is an operation that introduces slight random perturbations into latent vectors to enrich their diversity and prevent the population from converging to local optima too early. In the present study, we augment the diverse features in the user-guided vector by introducing random noise aligned with the

Algorithm 2 UIGAN for Personalized Music Synthesis.

```
Require: Rate of crossover, cro_rate; Rate of mutation, mut_rate;
    Number of music selected by the user, user_sel_count; Number
    of candidate music, cand_count; Trained MelGAN model,
    MelGAN; Max number of generation, max_gen.
Ensure: Synthesized music of a user, result;
for $i$ = 1 to cand_count do
    cand_music = Load_music(i);
end for
exit_flag = false;
init_cand_music = cand_music;
init_cand_mels = Audio_Mel(cand_music);
for $g$=1 to max_gen do
    selected_label, exit_flag, reset_flag =
        Inter_Player(cand_music, user_sel_count);
    if $g$ = max_gen+1 or exit_flag then;
        result = selected_music;
        break;
    end if
    if reset_flag then
        cand_music = init_cand_music;
        continue;
    end if
    mel_music = Audio_Mel(cand_music);
    selected_mels = mel_music(selected_label);
    crossed_mels = Crossover(selected_mels, cro_rate);
    mutated_mels = Mutation(crossed_mels, mut_rate);
    rand_num = cand_num - Count(selected_mels,
        crossed_mels, mutated_mels);
    rand_mels = Rand_Select(init_cand_mels, rand_num);
    cand_mels = List(mutated_mels, crossed_mels,
        selected_mels, random_mels);
    generated_music = MelGAN(cand_mels);
    denoising_music = DeNoise(generated_music);
    cand_music = denoising_music;
end for
```

distribution of the dataset. This approach establishes the foundation for the subsequent regeneration of the MelGAN model in the next step.

3.2.4 Regeneration. After a series of diverse optimization operations, the user-guided vectors are reused as input for the pre-trained MelGAN model, thereby facilitating the further regeneration of new candidate music. The user-guided vectors represent the user’s preferences, and the music generated from these vectors will achieve the subjective aesthetic goal.

3.2.5 Noise Reduction. Noise reduction constitutes the final phase of UIGAN. Given that random noise is introduced during the mutation stage in the user-guided latent vector, this study employs the spectral gating noise reduction method [38] to mitigate noise interference, thereby enhancing both sound quality and the overall user experience.

3.3 Pseudo Code

Algorithm 2 provides a detailed explanation of the pseudo-code and the key steps involved in UIGAN for the personalized music synthesis process. Initially, the *Load_music* function is responsible for importing the music that the user wants to adapt. The *Audio_Mel* function is utilized to convert music waveforms into mel-spectrograms. Next, the *Inter_Player* function plays a crucial role in obtaining user preferences for the candidate music through an interactive music player. This step involves variables like selected_label, exit_flag, and reset_flag to represent the selected music, exit flag, and reset flag, respectively.

Subsequently, the *Crossover* and *Mutation* functions are dedicated to performing the necessary crossover and mutation operations on the mel-spectrograms of the selected music, ensuring diversity in the user-guided latent vectors. Furthermore, the *Count* function tallies the number of user-guided latent vectors, and the remaining candidate vector is randomly selected using the *Rand_Select* function. The *List* function is employed to generate a list of latent vectors after music synthesis. The *MelGAN* function is responsible for generating candidate music from the latent vectors. Finally, the *DeNoise* function refines and denoises the synthesized music, ensuring the sound quality of the music for this generation.

4 EXPERIMENTAL ANALYSIS AND DISCUSSION

This section aims to validate the performance of our proposed model in music synthesis from quantitative and qualitative perspectives, including diversity and creativity analysis, along with subjective analysis. Additionally, the primary objective of this section is to illustrate the pivotal role of humans in synthesizing diverse and creative music. Subsequently, we will provide details on the dataset employed in this experiment, outline implementation specifics, and present the results of the analysis and discussion.

Table 1: Experimental parameters.

Parameter size	Value
Crossover method [4]	’Uniform crossover’
Crossover rate [4, 42]	50%
Number of music selected by the user	2
Number of candidate music [4, 52]	10
Max number of generation	20
Duration of time for synthesizing music [16][19]	4 seconds

4.1 Experimental Setting

4.1.1 Dataset. This experiment utilized the MAESTRO piano dataset [14][1], employing the WAV data format.

4.1.2 Implementations Details. The baseline methods employed in this experiment include WaveNet [48] [2], WaveGAN [16] [3], and

[1] https://magenta.tensorflow.org/datasets/maestro

[2] https://github.com/ibab/tensorflow-wavenet

[3] https://github.com/mostafaelaraby/wavegan-pytorch/tree/master

Table 2: Diversity and creativity analysis: We utilize the FAD, precision, recall, density, and coverage metrics to quantify the diversity and creativity of generated music, showcasing the role of human creativity in enhancing music diversity and creativity. P&R&D&C represents the averages of precision, recall, density, and coverage. Where ↑ and ↓ represent positive and negative trends, respectively. Bold values represent the best performance for each metric.

Dataset	Method	FAD ↓	Precision ↑	Recall ↑	Density ↑	Coverage ↑	P&R&D&C ↑
Mastro	WaveNet [48]	33.51	0.53	0.49	0.41	0.11	0.39
	WaveGAN [16]	24.58	0.72	**0.78**	0.48	0.17	0.54
	MelGAN [19]	24.92	**0.79**	0.62	**0.54**	0.15	0.53
	UIGAN	**22.14**	0.77	0.71	0.53	**0.21**	**0.56**

MelGAN [19] [4]. These techniques are well-acknowledged for their proficiency in synthesizing audio waveforms. For detailed parameter information and code references related to the baseline methods, please refer to the respective published papers. It is noteworthy that the MelGAN implementation used in UIGAN is consistent with the original paper.

Ten participants were invited to participate in the music synthesis and subjective evaluation analysis of related synthesized music. The experimental setting and user-friendly interactive interface (Figure. 3) facilitated participation without the need for professional music knowledge. Additionally, participants were given the flexibility to take short breaks during the experiment.

The parameters utilized in the present experiment are outlined in Table 1. The adoption of uniform crossover with a crossover rate of 50% is attributed to its excellent performance in image synthesis [4, 42]. Considering both user fatigue and the equal probability of each music's mel-spectrogram being synthesized into the user-guided latent vector, we set the number of user choices to 2. The mutation rate is set to 0.01, resembling the process of setting the learning rate in deep learning. A high mutation rate introduces excessive noise content, posing challenges for UIGAN convergence. The number of synthesized musical pieces selected by the user in the audio list is ten, including two pieces synthesized by crossed vectors, two by mutated vectors and selected mel-spectrograms, and the remaining four are randomly selected from the initial candidate set. The maximum generation parameter is set to 20. Additionally, we have introduced an optimization-stopping mechanism to ensure the experiment's completion while avoiding unnecessary iterations and user fatigue [49]. Finally, the duration of the synthesized music is set to four seconds [16, 19].

4.2 Diversity and Creativity Analysis and Discussion

4.2.1 Evaluation Metrics. Fréchet audio distance (FAD) [17], precision [20], recall [20], density [31], and coverage [31] are applied to evaluate the diversity and creativity of synthetic music on the MAESTRO dataset. These metrics are widely used to assess the creativity and diversity of image and audio generation models in recent years [16, 18–20, 31].

4.2.2 Experimental Results. Table 2 presents the diversity and creativity results of the generated music. We observe that UIGAN surpasses the baseline methods in terms of FAD, coverage, and P&R&D&C. Specifically, the proposed method achieves optimal values of 22.14, 0.21, and 0.56 in these three aspects, respectively. Moreover, UIGAN demonstrates comparable performance in various other metrics, including precision, recall, and density, with only a slight deviation from the highest values of these metrics. For example, MelGAN exhibits only a slight improvement of 0.02 in precision and 0.01 in density compared to UIGAN.

In contrast to the previous MelGAN model, the introduction of human interaction contributes to exploring various musical elements and combinations, thereby enhancing musical diversity. The excellent performance of UIGAN can be attributed to two factors. Firstly, the model combines the proficiency of MelGAN in scene restoration with the advantages of WaveGAN in enhancing diversity. Secondly, the model introduces the potential for greater diversity through interactive evolutionary manipulation while preserving the underlying musical scene depicted by the mel-spectrogram of music.

4.3 Subjective Analysis and Discussion

4.3.1 Evaluation Metrics. Subjective evaluation analysis is an essential component in assessing human-machine works, playing a crucial role for researchers in comprehending users' authentic experiences with the works and systems. The 5-point mean opinion score (MOS) was employed to gauge users' genuine perceptions of their synthesized music [16, 18, 19, 48]. Ten participants rated the four pieces of music based on their auditory preferences, with a score of 1 indicating the worst and 5 denoting the best. It is noteworthy that three of the musical pieces were pre-generated through various baseline methods, whereas one piece of music was generated through the UIGAN user interface.

The Wilcoxon signed-rank test is a non-parametric hypothesis test designed for data that does not adhere to a normal distribution, utilized to assess significant differences between matched samples. This testing approach has been effectively employed in text-to-speech synthesis [46] and speech synthesis [40]. In this study, the Wilcoxon signed-rank test with a p-value of 0.05 [46] is utilized to evaluate the significant differences between UIGAN and other music synthesis methods in 5-point MOS.

4.3.2 Experiemntal Results. Table 3 presents the 5-point MOS results from participants evaluating the synthesized music. We can observe that UIGAN achieved the highest MOS of 3.9 with a standard deviation (SD) of 0.7, while WaveNet had a minimum value of 1.7 with an SD of 0.64. It indicates that the proposed method can effectively meet users' requirements for music synthesis. At the same time, the p-values from Wilcoxon signed-rank test for both

[4] https://github.com/descriptinc/melgan-neurips

Table 3: Subjective evaluation analysis: The 5-point mean opinion score with standard deviation (±SD), provides insight into participants' genuine experiences. Additionally, the p-value of Wilcoxon signed-rank serves as an indicator of significant differences between synthetic music methods. We can observer that there is a significant difference comparing our proposed UIGAN with other three competitive generation models.

Dataset	Method	MOS	p-value
Mastro	UIGAN	**3.9 ± 0.70**	—
	WaveNet [48]	1.7 ± 0.64	0.006
	WaveGAN [16]	2.3 ± 1.10	0.017
	MelGAN [19]	3.3 ± 0.64	0.014

UIGAN and the other methods are all below 0.05, suggesting a notable distinction in MOS between UIGAN and these methods.

During the initial stages of interactive evolutionary manipulation, users' initial uncertainty or vague understanding of music may lead to cognitive disparities and uncertainties, posing a challenge in crafting works that authentically connect with users. Nevertheless, as user preferences progressively crystallize during continuous interaction, the vocoder can adeptly produce audio compositions that harmonize with user inclinations by acquiring latent vectors tailored to those preferences.

5 CONCLUSION

This paper introduces UIGAN, a user-guided generation framework for personalized music synthesis, which presents one of the originalites in this work. UIGAN leverages IEC's proficiency in capturing user preferences and GAN's capacity for autonomously generating high-quality music to enhance the credibility and diversity of generative AI in music synthesis, fostering computational artistic creativity. Validation of the proposed method's advantages in music synthesis includes diversity and creativity analysis, along with subjective evaluation. Furthermore, it confirms the pivotal role of humans in synthesizing diverse and creative music, establishing a co-creative paradigm for generative AI and human-machine interaction. The model empowers humans to guide and manipulate the potential space within deep learning models.

This study reveals insights into stimulating computational artistic creativity among humans and addresses potential issues in current generative AI practices. Future research directions include investigating UIGAN's performance in synthesizing diverse music with more datasets, encouraging comprehensive comparative experiments, analysing human auditory perception, enhancing search performance of IEC optimization algorithm and addressing the challenges posed by the complexities of music. Despite rapid advancements in other AI domains, such as computer vision and natural language processing, progress in music synthesis has been comparatively slower. Researchers need to tackle challenges related to musical knowledge and intricate temporal relationships among musical elements, such as notes, chords, and melodies, etc. Persistent concerns about user fatigue in IEC and human-machine interaction necessitate exploration of ways to minimize fatigue during interactions in the future. These considerations provide valuable insights and ideas for future endeavors in music synthesis and human-machine interaction.

ACKNOWLEDGMENTS

The authors extend their sincere appreciation to all participants who actively contributed to this study. The authors affirm that they have no conflicts of interest associated with this work. Informed consent was meticulously acquired from each individual participant involved in the study. All procedures conducted in studies involving human participants adhered rigorously to the ethical standards stipulated by the institutional and/or national research committee. Furthermore, these procedures were conducted in accordance with the principles outlined in the 1964 Helsinki Declaration and its subsequent amendments or comparable ethical standards.

REFERENCES

[1] Nuha Aldausari, Arcot Sowmya, Nadine Marcus, and Gelareh Mohammadi. 2022. Video generative adversarial networks: a review. *ACM Computing Surveys (CSUR)* 55, 2 (2022), 1–25. https://doi.org/10.1145/3487891

[2] Matthew Baas and Herman Kamper. 2023. GAN You Hear Me? Reclaiming Unconditional Speech Synthesis from Diffusion Models. In *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, Doha, Qatar, 906–911. https://doi.org/10.1109/SLT54892.2023.10023153

[3] David Baidoo-Anu and Leticia Owusu Ansah. 2023. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI* 7, 1 (2023), 52–62. https://doi.org/10.61969/jai.1337500

[4] Philip Bontrager, Wending Lin, Julian Togelius, and Sebastian Risi. 2018. Deep interactive evolution. In *Computational Intelligence in Music, Sound, Art and Design: 7th International Conference(EvoMUSART)*. Springer, California,USA, 267–282. https://doi.org/10.1007/978-3-319-77583-8_18

[5] Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. MusicLDM: Enhancing Novelty in Text-to-Music Generation Using Beat-Synchronous Mixup Strategies. https://doi.org/10.48550/arXiv.2308.01546 arXiv:2308.01546 [cs.SD]

[6] Sung-Bae Cho and Joo-Young Lee. 2002. A human-oriented image retrieval system using interactive genetic algorithm. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 32, 3 (2002), 452–458. https://doi.org/10.1109/TSMCA.2002.802812

[7] Grant Cooper. 2023. Examining science education in chatgpt: An exploratory study of generative artificial intelligence. *Journal of Science Education and Technology* 32, 3 (2023), 444–452. https://doi.org/10.1007/s10956-023-10039-y

[8] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A Generative Model for Music. https://doi.org/10.48550/arXiv.2005.00341 arXiv:2005.00341 [eess.AS]

[9] Chris Donahue, Julian McAuley, and Miller Puckette. 2019. Adversarial Audio Synthesis. https://doi.org/10.48550/arXiv.1802.04208 arXiv:1802.04208 [cs.SD]

[10] Malachy Eaton. 2013. An approach to the synthesis of humanoid robot dance using non-interactive evolutionary techniques. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, Manchester, UK, 3305–3309. https://doi.org/10.1109/SMC.2013.563

[11] Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R Frank, Matthew Groh, Laura Herman, Neil Leach, et al. 2023. Art and the science of generative AI. *Science* 380, 6650 (2023), 1110–1111. https://doi.org/10.1126/science.adh4451

[12] Gunther Eysenbach et al. 2023. The role of ChatGPT, generative language models, and artificial intelligence in medical education: a conversation with ChatGPT and a call for papers. *JMIR Medical Education* 9, 1 (2023), e46885. https://doi.org/10.2196/46885

[13] Curtis Hawthorne, Ian Simon, Adam Roberts, Neil Zeghidour, Josh Gardner, Ethan Manilow, and Jesse Engel. 2022. Multi-instrument Music Synthesis with Spectrogram Diffusion. https://doi.org/10.48550/arXiv.2206.05408 arXiv:2206.05408 [cs.SD]

[14] 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. https://doi.org/10.48550/arXiv.1810.12247 arXiv:1810.12247 [cs.SD]

[15] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE, Honolulu,HI, 1125–1134. https://doi.org/10.1109/CVPR.2017.632

[16] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaron Oord, Sander Dieleman, and Koray Kavukcuoglu. 2018. Efficient neural audio synthesis. In *International Conference on Machine Learning*. PMLR, Stockholm, Sweden, 2410–2419. https://doi.org/

10.48550/arXiv.1802.08435

[17] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. 2019. Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms. https://doi.org/10.48550/arXiv.1812.08466 arXiv:1812.08466 [eess.AS]

[18] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2021. DiffWave: A Versatile Diffusion Model for Audio Synthesis. https://doi.org/10.48550/arXiv.2009.09761 arXiv:2009.09761 [eess.AS]

[19] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems* 32 (2019), 1–12. https://doi.org/10.48550/arXiv.1910.06711

[20] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2019. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems* 32 (2019), 1–10. https://doi.org/10.48550/arXiv.1904.06991

[21] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. 2016. Autoencoding beyond pixels using a learned similarity metric. In *International conference on machine learning*. ACM, New York, 1558–1566. https://doi.org/10.48550/arXiv.1512.09300

[22] Jae Hyun Lim and Jong Chul Ye. 2017. Geometric GAN. https://doi.org/10.48550/arXiv.1705.02894 arXiv:1705.02894 [stat.ML]

[23] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. 2022. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI conference on artificial intelligence*. AAAI, Virtual Event, 11020–11028. https://doi.org/10.1609/aaai.v36i10.21350

[24] Jen-Yu Liu, Yu-Hua Chen, Yin-Cheng Yeh, and Yi-Hsuan Yang. 2020. Unconditional audio generation with generative adversarial networks and cycle regularization. https://doi.org/10.48550/arXiv.2005.08526 arXiv:2005.08526 [cs.SD]

[25] Roisin Loughran and Michael O'Neill. 2020. Evolutionary music: applying evolutionary computation to the art of creating music. *Genetic Programming and Evolvable Machines* 21 (2020), 55–85. https://doi.org/10.1007/s10710-020-09380-7

[26] Shihan Lu, Mianlun Zheng, Matthew C Fontaine, Stefanos Nikolaidis, and Heather Culbertson. 2022. Preference-driven texture modeling through interactive generation and search. *IEEE Transactions on Haptics* 15, 3 (2022), 508–520. https://doi.org/10.1109/TOH.2022.3173935

[27] Janos Madar, Janos Abonyi, and Ferenc Szeifert. 2005. Interactive particle swarm optimization. In *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*. IEEE, Wroclaw, Poland, 314–319. https://doi.org/10.1109/ISDA.2005.58

[28] Dwilya Makiwan, Kaori Yoshida, and Mario Koppen. 2017. Interactive evolutionary computation of color palette design enhanced by impression words. In *2017 International Conference on Platform Technology and Service (PlatCon)*. IEEE, Busan, South Korea, 1–6. https://doi.org/10.1109/PlatCon.2017.7883712

[29] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. 2017. SampleRNN: An unconditional end-to-end neural audio generation model. https://doi.org/10.48550/arXiv.1612.07837 arXiv:1612.07837 [cs.SD]

[30] Michael Muller, Lydia B Chilton, Anna Kantosalo, Q Vera Liao, Mary Lou Maher, Charles Patrick Martin, and Greg Walsh. 2023. GenAICHI 2023: Generative AI and HCI at CHI 2023. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg, Germany, 1–7. https://doi.org/10.1145/3544549.3573794

[31] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. 2020. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*. PMLR, Virtual Event, 7176–7185. https://doi.org/10.48550/arXiv.2002.09797

[32] Xingang Pan, Ayush Tewari, Thomas Leimkühler, Lingjie Liu, Abhimitra Meka, and Christian Theobalt. 2023. Drag your gan: Interactive point-based manipulation on the generative image manifold. In *ACM SIGGRAPH 2023 Conference Proceedings*. ACM, New York, 1–11. https://doi.org/10.1145/3588432.3591500

[33] Wenjun Pang and KC Hui. 2010. Interactive evolutionary 3d fractal modeling. *The Visual Computer* 26 (2010), 1467–1483. https://doi.org/10.1007/s00371-010-0500-8

[34] Yan Pei. 2023. A comprehensive and brief survey on interactive evolutionary computation in sound and music composition for algorithmic auditory and acoustic design with human-in-the-loop. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*. ACM, Lisbon, Portugal, 1990–1996. https://doi.org/10.1145/3583133.3596301

[35] Yan Pei and Hideyuki Takagi. 2013. Triple and quadruple comparison-based interactive differential evolution and differential evolution. In *Proceedings of the twelfth workshop on Foundations of genetic algorithms XII*. Springer, New York, 173–182. https://doi.org/10.1145/2460239.2460255

[36] Hua Peng, Huosheng Hu, Fei Chao, Changle Zhou, and Jing Li. 2016. Autonomous robotic choreography creation via semi-interactive evolutionary computation. *International Journal of Social Robotics* 8 (2016), 649–661. https://doi.org/10.1007/s12369-016-0355-x

[37] Juan C Quiroz, Sushil J Louis, Anil Shankar, and Sergiu M Dascalu. 2007. Interactive genetic algorithms for user interface design. In *2007 IEEE congress on evolutionary computation*. IEEE, Singapore, 1366–1373. https://doi.org/10.1109/CEC.2007.4424630

[38] Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. 2020. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS computational biology* 16, 10 (2020), e1008228. https://doi.org/10.1371/journal.pcbi.1008228

[39] Tim Salimans and Durk P Kingma. 2016. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems* 29 (2016), 1–11. https://doi.org/10.48550/arXiv.1602.07868

[40] Kai Shigemi, Shuji Komeiji, Takumi Mitsuhashi, Yasushi Iimura, Hiroharu Suzuki, Hidenori Sugano, Koichi Shinoda, Kohei Yatabe, and Toshihisa Tanaka. 2023. Synthesizing speech from ecog with a combination of transformer-based encoder and neural vocoder. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Rhodes Island, Greece, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10097004

[41] Xiaoyan Sun, Dunwei Gong, Yaochu Jin, and Shanshan Chen. 2013. A new surrogate-assisted interactive genetic algorithm with weighted semisupervised learning. *IEEE Transactions on Cybernetics* 43, 2 (2013), 685–698. https://doi.org/10.1109/TSMCB.2012.2214382

[42] Gilbert Syswerda et al. 1989. Uniform crossover in genetic algorithms.. In *International Computer Games Association*, Vol. 3. IOS Press, London, 2–9. https://dl.acm.org/doi/abs/10.5555/645512.657265

[43] Koray Tahiroglu, Miranda Kastemaa, and Oskar Koli. 2021. Ganspacesynth: A hybrid generative adversarial network architecture for organising the latent space using a dimensionality reduction for real-time audio synthesis. In *Proceedings of the 2nd Joint Conference on AI Music Creativity*. AIMC, Virtual Event, 1–11. https://doi.org/10.5281/zenodo.5137902

[44] Hideyuki Takagi and Miho Ohsaki. 2007. Interactive evolutionary computation-based hearing aid fitting. *IEEE Transactions on Evolutionary Computation* 11, 3 (2007), 414–427. https://doi.org/10.1109/TEVC.2006.883465

[45] Hideyuki Takagi and Denis Pallez. 2009. Paired comparison-based interactive differential evolution. In *2009 World Congress on Nature & Biologically Inspired Computing (NaBIC)*. IEEE, Coimbatore, India, 475–480. https://doi.org/10.1109/NABIC.2009.5393359

[46] Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. 2024. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024). https://doi.org/10.1109/TPAMI.2024.3356232

[47] Nao Tokui and Hitoshi Iba. 2000. Music composition with interactive evolutionary computation. In *Proceedings of the third international conference on generative art*. ACM, Milan,Italy, 215–226. https://cir.nii.ac.jp/crid/1570854175183893888

[48] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. https://doi.org/10.48550/arXiv.1609.03499 arXiv:1609.03499 [cs.SD]

[49] Yanan Wang, Yan Pei, Shindo Hayato, Qing Liu, and Hai-Peng Ren. 2023. An Interactive Differential Evolution Method with Human Auditory Perception for Sound Composition. *IEEE Transactions on Cognitive and Developmental Systems* (2023), 1–13. https://doi.org/10.1109/TCDS.2023.3339193

[50] Somlak Wannarumon, Erik LJ Bohez, and Kittinan Annanon. 2008. Aesthetic evolutionary algorithm for fractal-based user-centered jewelry design. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* 22, 1 (2008), 19–39. https://doi.org/10.1017/S0890060408000024

[51] Lonce Wyse, Purnima Kamath, and Chitralekha Gupta. 2022. Sound Model Factory: An Integrated System Architecture for Generative Audio Modelling. In *International Conference on Computational Intelligence in Music, Sound, Art and Design (Part of EvoStar)*. Springer, Madrid, Spain, 308–322. https://doi.org/10.1007/978-3-031-03789-4_20

[52] Nicola Zaltron, Luisa Zurlo, and Sebastian Risi. 2020. Cg-gan: An interactive evolutionary gan-based approach for facial composite generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. AAAI, New York, NY, USA, 2544–2551. https://doi.org/10.1609/aaai.v34i03.5637

[53] Ning Zhang, Ruru Pan, Lei Wang, Yang Wu, and Weidong Gao. 2020. Pattern design and optimization of yarn-dyed plaid fabric using modified interactive genetic algorithm. *The Journal of The Textile Institute* 111, 11 (2020), 1652–1661. https://doi.org/10.1080/00405000.2020.1738617