A Novel Probabilistic Diffusion Model Based on the Weak Selection Mimicry Theory for the Generation of Hypnotic Songs

Wenkai Huang * and Feng Zhan

School of Mechanical and Electrical Engineering, Guangzhou University, Guangzhou 510006, China; 2007300008@e.gzhu.edu.cn
* Correspondence: smallkat@gzhu.edu.cn

Abstract: The constraints in traditional music style transfer algorithms are difficult to control, thereby making it challenging to balance the diversity and quality of the generated music. This paper proposes a novel weak selection–based music generation algorithm that aims to enhance both the quality and the diversity of conditionally generated traditional diffusion model audio, and the proposed algorithm is applied to generate natural sleep music. In the inference generation process of natural sleep music, the evolutionary state is determined by evaluating the evolutionary factors in each iteration, while limiting the potential range of evolutionary rates of weak selection–based traits to increase the diversity of sleep music. Subjective and objective evaluation results reveal that the natural sleep music generated by the proposed algorithm has a more significant hypnotic effect than general sleep music and conforms to the rules of human hypnosis physiological characteristics.

Keywords: probabilistic diffusion model; weak selection mimicry theory; hypnotic songs; sleep music

MSC: 68T99

1. Introduction

Sleep is crucial to human health and happiness [1,2], but many people occasionally or consistently experience sleep disorders and disruptions [3,4], which significantly affect their quality of life [5]. Natural sleep music [6,7] is an effective means to alleviate sleep disorders and is widely used in clinical treatment and therapeutic applications [8,9]. Listening to natural sleep music helps distract the listener from worries, increases relaxation levels, and induces deep sleep by increasing slow–wave activity [10]. Therefore, creating novel and suitable natural sleep music is essential for relieving sleep disorders and clinical medicine [11,12] by alleviating anxiety and stress, relieving pain, and bolstering the immune system. In recent years, using deep learning to create hypnotic music has become a popular research area [13–16].

Currently, white Gaussian noise is commonly used in popular hypnosis music generation [17]. However, Alvarsson et al. [18] found that 95% of participants felt tired and unhappy when listening to white noise, which contradicts the purpose of hypnosis. In contrast, sleep music is described as inducing comfort and pleasant relaxation, which has a positive effect on sleep. A study was conducted by researchers [19] in response to this issue, which compared the unique tracks between the sleep playlist dataset (SPD) and the music streaming session dataset (MSSD) using Welch’s t-test, and the comparative results are presented in Figure 1. Furthermore, the details of the indicators are presented in Appendix A. This study found that the most significant difference between the two was in terms of loudness, followed by energy, sonority, and instrumentality. Therefore, when

Citation: Huang, W.; Zhan, F. A Novel Probabilistic Diffusion Model Based on the Weak Selection Mimicry Theory for the Generation of Hypnotic Songs. Mathematics 2023, 11, 3345. https://doi.org/10.3390/math11153345
Academic Editor: Faheem Sufi
Received: 21 June 2023; Revised: 25 July 2023; Accepted: 27 July 2023; Published: 30 July 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).
Creating hypnosis music, it is important to avoid using overly stimulating and unpleasant music elements, such as soft sounds, gentle melodies, and natural sounds that promote relaxation and tranquility, and avoid overly stimulating and unpleasant sound elements, such as dissonant notes and sharp, piercing sounds.

![Figure 1. Comparison of audio functions between sleep music and general music. The figure presents individual audio features represented by smoothed density plots and box plots below, with vertical lines indicating medians. The corresponding Cohen’s d values are used to compare sleep music and general music.](image)

Therefore, numerous studies have attempted to design or optimize neural network models based on deep learning to generate efficient hypnotic music [20]. According to the literature review of existing deep learning–based music generation and style transfer, the creation of hypnotic music is mainly based on two kinds of music style transfer models: generative adversarial network (GAN)–based methods [21–23] and variational autoencoder (VAE) [22].

Deep music generation involves using computer systems that employ deep learning network architectures to automatically generate music [24]. Within the realm of music generation research, deep learning algorithms have emerged as the predominant method [14]. A significant advancement in this field was made by Donahue et al. [23], who were the first to explore the application of generative adversarial networks (GANs) for unsupervised raw audio synthesis. Expanding on this progress, Engel et al. [26] introduced GANSynth, a model that utilizes GANs to generate high–fidelity and locally coherent audio by effectively modeling log magnitudes and instantaneous frequencies with precise frequency resolution in the spectral domain. While GANs have proven to be powerful, they are also known for their challenging nature of training and their limited applicability to sequential data [14].

On the basis of the literature review of music generation using GANs [27,28], it has been found that GAN–based methods face some challenges in the field of music generation, including embedded control and interactivity. Currently, the commonly used
CycleGAN in this field may result in the generation of music that is not harmonious due to information loss when using perceptually enriched spectrograms for inverse deduction [29]. Additionally, certain researchers have also used conditional GANs to achieve more finely controlled music, but the resulting music continues to have unnatural issues [30,31]. In the field of VAE, Hu et al. [22] used the VAE framework with only two input music pieces for style transfer to generate hypnotic music. However, the generated music is often too similar to the style of input samples, thereby lacking novelty [32]. Because it is necessary to consider multiple factors such as melody and rhythm for hypnotic music, the structure of the music is hierarchical and multimodal, with multilevel composition features (pitch and rhythm) and low-level features (sound texture and timbre) [14,16,33].

Traditional music style transfer methods [34,35] can be used to create natural sleep music, but there are some limitations with existing methods. Natural sleep music is often composed of multiple layers of instruments and genres [26], which can make it challenging to isolate and manipulate individual elements of the music during the style transfer process. First, the types of natural audio that users are interested in are limited [36], and it is difficult for existing methods to generate specific styles of hypnotic music according to user needs. Second, existing methods have difficulty adjusting the constraint conditions on music styles [34], thereby resulting in the generated music being rather random [37] and making it difficult to generate music that meets user requirements. In addition, some strongly constrained algorithms [38] may lack creative diversity in the generated music and fail to explore dynamic changes in pieces that are actually creative. In general, specific constraints and controlled conditions of constraint algorithms pose significant challenges for generation models. Therefore, an adaptive constraint model that can learn multilevel music features is needed to improve the reliability and adaptability of natural sleep music generation [39,40].

In cutting-edge research [13,41,42], the conditional diffusion probabilistic model has gradually emerged as one of the most commonly used techniques in deep learning–based music creation models. The conditional diffusion probabilistic model consists of two processes: a diffusion process and an inverse process. The diffusion process is a Markov chain with fixed parameters that gradually transforms complex data into isotropic Gaussian distributions by adding Gaussian noise; the inverse process is a Markov chain implemented by neural networks that iteratively recovers the original data from white Gaussian noise [43]. Although the generation task of natural sleep music can be viewed as a traditional conditional diffusion model framework problem, the music generated by this model lacks overall sound quality and clear details, thereby resulting in a lack of fluidity. Therefore, in order to generate hypnotic music that is more in line with human physiology and music theory, we need to improve the conditional constraint method of the traditional diffusion model to adjust the capacity of learning natural audio from the source audio and create natural hypnotic music. Inspired by the adaptive weak selection mimetic dynamics [44,45], in complex natural selection, the phenotypic response of mimetic samples to the evolutionary rate and direction of multiple traits results in the final evolution of a diverse phenotype adapted to the environment [46]. In music creation, this idea implies preserving the structural information of the source audio (melody), while integrating the low-level attributes (rhythm and frequency) corresponding to natural audio to generate hypnotic music that is more in line with human physiology and music theory. On the basis of these starting points, the weak selection mimetic algorithm is introduced into the diffusion model.

The weak selection mimetic algorithm imitates the weak selection approximation of species to multiple perceived environmental components [47,48], which restricts the choice of better imitators of a single target. The weak selection imitation [44] assumes that the change of each allele is considered to be a continuous random variable that follows a distribution governed by the diffusion equation. With the driving force of genetic variation, an exact time–dependent solution for the diffusion equation of the selectively neutral population can be obtained [49–51]. Like perfect imitation, weak selection imitation is also
characterized by adaptability, where local adaptability is stronger than global adaptability. This phenomenon may reflect certain restrictions on signal production, which (at least in certain cases) hinder individuals from reacting to better choices [52]. Similarly, in the specific implementation of the creation of natural sleep music, a general music approach is adopted to selectively imitate the multilevel features of natural audio in order to dynamically enhance conditional features [53]. In summary, in order to solve the problems of conditional constraints and insufficient global information in the conditional diffusion model [54,55], we used a conditional module to obtain the overall characteristics of natural audio and used it as a local embedding input for the weakly selective mimetic algorithm. Subsequently, we used this algorithm to calculate the rationality between the diffusion sample and the local embedding, performed weak selection approximation, obtained the diffusion sample at the approximate equilibrium, and finally transmitted it to the corresponding hierarchical structure of the denoising network. By optimizing the evidence lower bound of the data distribution [56], it can be effectively trained without adversarial feedback [57], thereby leading to the generation of natural hypnosis music waveforms that match the ground truth distribution. According to subjective and objective evaluation by the testers, our generated natural hypnosis music is more efficient and can help testers fall asleep quickly. The following points summarize the main contribution of this paper:

(1) A novel weakly selective mimetic music generation algorithm is proposed for the generation of natural hypnosis music. The proposed algorithm enables the adaptive control of the model’s ability to learn global and local features by calculating the conditional evolutionary inertia in the conditional diffusion probability model. Thus, a balance can be achieved between the diversity and quality of the generated natural hypnosis music.

(2) To accelerate the convergence speed of the weak selection mimetic algorithm and meet the feature requirements of hypnosis music, a novel conditional module is introduced for the proposed algorithm. This module can use features with multiple scales to generate local embeddings of the conditional audio in a low-cost and robust manner, thereby eliminating high-frequency components.

(3) A comprehensive evaluation of the algorithm is conducted through subjective and objective experiments. The generated music rhythm conforms to the physiological characteristics of human sleep and effectively improves the sleep efficiency of patients with sleep disorders.

2. Related Work

This article mainly involves the diffusion probability model and weak selection mimicry. Therefore, the principles of the conditional diffusion probability model and weak selection mimicry are briefly introduced in this section.

2.1. Conditional Diffusion Probability Model

In this section, the principles of the conditional diffusion probability model [55,58,59] are introduced. The process of the conditional diffusion probability model is mainly divided into two parts: (1) transforming the original data distribution to a Gaussian distribution through the diffusion process; (2) recovering data from Gaussian white noise through the reverse process based on a conditional regulator. These processes are depicted in Figure 2.
The basic principle underlying the diffusion process is to introduce the diffusion equation. In the diffusion process, each step of transfer is completed by a determined transfer operator, and the parameters of the transfer operator are obtained through training and learning. The diffusion process is sampled from the initial state data distribution $q(x_0)$, where $x_0 \sim q(x)$. Thereafter, it undergoes the Markov chain diffusion process of $T$ steps; thus, $x_0$ is transformed into $x_T$, where $x_T \sim q(x_T)$, and $q(x_T)$ conform to the Gaussian distribution.

$$q(x_T) := \int_{\mathbb{R}^T} q(x_0) \prod_{t=1}^{T} q(x_t | x_{t-1}) dx_0 \cdots dx_{T-1}. \quad (1)$$

In each step of the diffusion process $t$, where $t \in [1, T]$, a Gaussian noise of different levels will be added to $x_{t-1}$ to obtain $x_t$, according to a noise variance adjustment table $\beta = \{\beta_1, \ldots, \beta_T\}$.

$$q(x_t | x_{t-1}) := \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I). \quad (2)$$

If appropriate values of $\beta_t$ are used in the diffusion process, along with a sufficiently large $T$, $q(x_t)$ will approach an isotropic Gaussian distribution. Furthermore, there is a special property of diffusion process that $q(x_t|x_0)$ can be calculated in closed form in $O(1)$ time.

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\alpha_t} x_0, 1 - (1 - \alpha_t)^1), \quad (3)$$

where $\alpha_t$ can be described as

$$\alpha_t := \prod_{s=1}^{t} \alpha_s, \alpha_t = 1 - \beta_t. \quad (4)$$

The reverse process of the model is a Markov chain with learnable parameters $\theta$. The inverse process uses the Mel spectrogram, $y$, as a regulator. Since it is difficult to achieve an accurate inverse transformation distribution $q(x_{t-1}|x_t)$, we use a neural network with fixed parameters, $\theta$, to approximate this distribution ($\theta$ is a shared parameter in each step $t$).

$$p_\theta(x_{t-1}|x_t, y) := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, y, \sigma^2_t)). \quad (5)$$

Thus, the complete inverse process can be described in the following manner:
\[
p_\theta(x_0, \cdots, x_{T-1}|x_T, y) = \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, y),
\]
\[(6)\]

where \( p_\theta := \mathcal{N}(x_{t-1}; \mu_\theta(x_t, y), \beta_t I) \).

\[
\tilde{\mu}_t(x_t, y, t) := \frac{1}{\sqrt{\alpha_t}} \left( x_t + \frac{\beta_t}{\sqrt{1 - \alpha_t}} \epsilon_\theta(x_t, y, t) \right),
\]
\[(8)\]

\[
\tilde{\beta}_t := \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t} \beta_t.
\]
\[(9)\]

By reparametrizing Equation (3), we obtain
\[
x_t(x_0, \epsilon) = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \epsilon.
\]
\[(10)\]

Therefore, the objective function can be simplified in the following manner:
\[
E_{x_0, \epsilon} \left[ \frac{\beta_t^2}{2\bar{\alpha}_t^2(1 - \bar{\alpha}_t)} \| \epsilon - \epsilon_\theta(\sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \epsilon, y, t) \|^2 \right].
\]
\[(11)\]

The noise \( \epsilon \) added to \( x_0 \) is predicted and denoised through the neural network \( \epsilon_\theta \). We set \( \sigma^2 \) to \( \beta_t \), sample \( \epsilon \sim \mathcal{N}(0,1) \), and output \( \epsilon_\theta(\cdot) \) from the neural network. A sample \( x_T \) is drawn from \( x_T \sim \mathcal{N}(0, I) \), and the data sample distribution is obtained by running the inverse diffusion process through the neural network.

2.2. Theory of Weak Selection Mimicry

Mimicry is an evolutionary strategy used by organisms to imitate the appearance, sound, behavior, and other traits of other species to obtain a survival advantage [60]. For example, in the realm of sound, birds and insects are known to mimic the songs or calls of other species to attract mates or establish their territory [61,62]. Weak selection mimicry is a special case of mimicry in which the mimicry sample does not perfectly resemble its target but still exhibits a few of the target’s local traits [47,63]. Weak selection mimicry can actually represent a stable evolutionary outcome and is a common phenomenon in nature. During the process of natural selection, the mimicry sample undergoes genetic variation to reduce the difference between its own phenotype and the various perceptual components of the target, thereby evading predation. After undergoing long–term evolution, organisms eventually evolve to achieve the highest similarity to their target. In previous studies [64,65], natural selection was shown to regulate the direction of genetic evolution to facilitate adaptive coevolution of mimicry samples with the perceptual signals of their environment. Mimicry samples can adjust their phenotype to be similar to their target under various conditions, thereby aiming to achieve optimal similarity with the target. Compared to perfect mimicry, weak selection mimicry results in more adaptive mimicry samples that evolve to achieve the best similarity with their target and determine the optimal evasion strategy to obtain the maximum benefits (i.e., the lowest predator cost) [44].

3. Methodology

3.1. Problem Definition

The diffusion model is a generative model that includes two stages: the forward diffusion stage and the reverse diffusion stage. The forward diffusion stage involves the diffusion of the acoustic characteristic \( x_0 \) of an audio signal, thereby resulting in \( x_T \sim \mathcal{N}(0, I) \). The reverse diffusion stage, which belongs to the diffusion model, generates a conditional audio signal by using a process of denoising. The reverse network \( \epsilon_\theta \) reconstructs the original signal.
\[ \epsilon_\theta(x_t, R, t) = \mathcal{D}(E^y_t, E^x_t, t). \] (12)

In this model, \( \theta \) represents the parameters of the reverse process, \( R \) is the natural input audio, \( t \) is the timestep, and \( x_t \) is the diffusion sample at the current step. \( y \) is a continuous feature embedding, which is encoded as a one-hot vector \( E^y_t \) and integrated with the time step index \( t \) to form an additional guidance signal \( E^y_t \). This is combined with the decoder feature \( E^x_t \) and input into the decoder \( \mathcal{D} \) to reconstruct the controllable hypnotic music feature \( x_0 \).

The network is trained by maximizing the evidence lower bound (ELBO) of the data. It is worth noting that only a certain proportion of the latent variables (the active subset) encodes useful information, while the remaining proportion (the passive subset) only encodes \( x_t \). Therefore, the useful information is constrained to the active subset. The latent statistical parameters are inferred from the input condition. Since the dimension of the \( x_t \) component is fixed, the network relies on the conditional information embedded in the input condition module to minimize the reconstruction loss and generate the conditional audio waveform through the decoder. However, the direct addition of the conditional information as a prior can result in additional high-frequency noise, which is detrimental to audio quality. Therefore, a new strategy is required to constrain the high-frequency components in the process of learning conditional features and reduce the impact of mismatched noise on audio quality.

3.2. Weakly Selective Mimetic Music Generation Algorithm

Inspired by the evolution of biological mimicry [66,67], weakly selective mimetic exploration explores the differences between source distributions and multiple target distributions and allows the distribution of multiple mimetic traits to weakly evolve toward the optimal phenotype that adapts to the environment [68]. Similarly, considering that the features of a piece of music can be abstracted into different levels, this team utilized a conditional module to extract the global and local structure of the spectrogram of natural audio as a conditional embedding to discover and represent the hidden structure of natural music from natural audio. Furthermore, they used an alignment tool to extract the valence, arousal, rhythm, pitch, loudness, and timbre features induced by music as conditional information for the input of the diffusion model decoder. This information was then entered into the weakly selective mimetic music generation algorithm. The algorithm can constrain the high-frequency components in the process of learning conditional features and dynamically balance the inconsistency between the distribution of conditional embeddings and the diffusion samples.

3.2.1. Algorithm Framework

This section presents the weakly selective mimetic music generation algorithm with implementation details. The overall framework is illustrated in Figure 3. The algorithm is explained in detail below.
3.2.2. Evolution Status

In this context, the original audio and natural audio are respectively understood as the pseudo-samples $X$ and target $Y$ in order to evolve them into their best-performing forms $\Theta_x$ and $\Theta_y$. During the stable selection process, we calculate the similarity and distance between each attribute of the original audio and the attributes of the natural audio in the neighborhood to obtain the fitness function [69].

$$
\tilde{w}_x(x) = \exp \left[ -\tilde{g}_x(x - \Theta_x)^2 \right].
$$

$$
\tilde{w}_y(y) = \exp \left[ -\tilde{g}_y(y - \Theta_y)^2 \right].
$$

The strength of self-selection is determined by positive parameters $\hat{g}_x$ and $\hat{g}_y$.

Fitness dynamics is a mathematical model of phenotype approximation [42], used to describe the dynamic changes of fitness during the interactions and evolution of music. This model considers the fitness of two audio signals as the result of their interaction and, on the basis of the advantageous response received by the recipient, obtains the fitness under different music characteristics:

$$
\tilde{w}_x(x) = \sum_y \exp \left[ -\tilde{g}_x(x - y)^2 \right] f_y(y).
$$

$$
\tilde{w}_y(y) = \sum_x \exp \left[ -\tilde{g}_y(y - x)^2 \right] f_x(x).
$$
Lastly, it is assumed that the two fitness components act multiplicatively. This is a reasonable assumption, as the corresponding selection pressures act simultaneously at different timepoints. Therefore, the overall fitness function is \( \hat{w}_x(x) = w_x^\gamma(x) \cdot \hat{w}_x^\gamma(x) \), and \( \hat{w}_y = w_y^\gamma(y) \cdot \hat{w}_y^\gamma(y) \). The average fitness can be viewed as a dynamic variable determined by the subjective evaluation factor \( \zeta \) selected by humans, as well as the values of the music’s own selection intensity, \( \hat{\theta}_x \), and the selection intensity between music pieces, \( \hat{\theta}_y \).

### 3.2.3. Mutation Strategy

Under the influence of natural selection, gene–based mutations and the gradual effects of natural selection are the long–term mechanisms for the potential evolution of ecosystems [70]. Evolutionary dynamics can accurately describe the evolution of driving audio diffusion. Specifically, evolutionary factors, including the strength of self–selection of the source audio and the selection intensity between the source and target audio, determine the characteristics of multiple traits in both source and target audio. The introduction of evolutionary factors enables the identification of the evolutionary state and the evaluation of the probability of evolution for each iteration. The following automatic adjustment learning strategy function fully describes the evolutionary dynamics:

\[
\frac{d\Psi_i}{dt} = \Psi_i(1 - \Psi_i) \frac{\partial \hat{w}_x}{\partial \Psi_i}
\]

\[
\frac{dY_j}{dt} = Y_j(1 - Y_j) \frac{\partial \hat{w}_y}{\partial Y_j}.
\]

The symbols \( \Psi_i \) and \( Y_j \) represent the gene frequencies of the \( i \)-th simulated trait and the \( j \)-th target trait, respectively. Learning global features is advantaged by a larger evolutionary inertia, while learning local features is benefitted by a smaller evolutionary inertia. When the Gaussian functions in Equations (13) and (14) are approximated with quadratics, and when terms of quadratic and higher order in both \( \hat{\theta} \) and \( \hat{\phi} \) are neglected, the average fitness of the two traits, denoted by \( \hat{w}_x \) and \( \hat{w}_y \), can be calculated. The calculation of these variables is simplified, leading to the following result:

\[
\hat{w}_x(x) = 1 - (\hat{\theta}_x + \hat{\phi}_x) \left\{ \left[ x - \hat{\theta}_x(\bar{y}) \right]^2 + G_x + G_y + \ldots \right\},
\]

\[
\hat{w}_y(y) = 1 - (\hat{\theta}_y + \hat{\phi}_y) \left\{ \left[ y - \hat{\theta}_y(x) \right]^2 + G_x + G_y + \ldots \right\}.
\]

The average feature of the source audio X is represented in the following manner:

\[
\bar{x} = x_m + 2 \sum_i a_i (\Psi_i - 1/2),
\]

where the genetic variance is represented as \( G_x = 2 \sum_i a_i^2 \Psi_i (1 - \Psi_i) \). Here, \( a_i \) represents the impact factor associated with the \( i \)-th genetic locus. Similarly, the average feature and genetic variance of the target y are represented by the following analogous expressions: \( \bar{y} = y_m + 2 \sum_j \beta_j (Y_j - 1/2) \), \( G_y = 2 \sum_j \beta_j^2 Y_j (1 - Y_j) \). Here, \( x_m \) and \( y_m \) represent the intermediate values of the x and y traits, respectively.

\[
\hat{\theta}_x(\bar{y}) = \theta_x + \frac{\hat{\theta}_x}{\hat{\theta}_x + \hat{\phi}_x} (\bar{y} - \theta_x),
\]

\[
\hat{\theta}_y(\bar{x}) = \theta_y + \frac{\hat{\theta}_y}{\hat{\theta}_y + \hat{\phi}_y} (\bar{x} - \theta_y).
\]

The respective expressions represent the optimal traits that the traits x and y attain at the average fitness under weak selection. By substituting Equations (19, 20) and (21, 22) into Equation (17, 18), Equation (23, 24) can be obtained, thereby formulating the probability of the coordinated evolution of traits under weak selection.
\[
\frac{d\Psi_i}{dt} = \Psi_i(1 - \Psi_i)\alpha_i^2\left(\bar{\theta}_x + \bar{\theta}_y\right)\left[2\Psi_i - 1 - \frac{\bar{\theta}_x(\bar{\gamma})}{\alpha_i}\right],
\]
(23)

\[
\frac{d\Psi_j}{dt} = Y_j(1 - Y_j)\beta_j^2\left(\bar{\theta}_x + \bar{\theta}_y\right)\left[2Y_j - 1 - \frac{\bar{\theta}_y(\bar{\gamma})}{\beta_j}\right].
\]
(24)

3.2.4. Updating of Genetic Variation

On the basis of the interaction between the audio and the change in the average phenotype, differential equations can be used to explain the evolution of the mimetic samples and the target phenotype to achieve the evolution of mimetic samples under weak selection, and the interplay between interspecific interaction and the effects of natural selection can be incorporated into the evolutionary model.

\[
\frac{d\bar{x}}{dt} = 2G_x(\bar{\theta}_x + \bar{\theta}_y)[\bar{\theta}_x(\bar{\gamma}) - \bar{x}] - (\bar{\theta}_x + \bar{\theta}_y)M_{ix},
\]
(25)

\[
\frac{d\bar{y}}{dt} = 2G_y(\bar{\theta}_x + \bar{\theta}_y)[\bar{\theta}_y(\bar{\gamma}) - \bar{y}] - (\bar{\theta}_x + \bar{\theta}_y)M_{iy},
\]
(26)

where the \(M_i\) values are the \(i\)-th central moments of the phenotypic distribution, which measure asymmetry [69]. Both the genetic variances and the central moments of the phenotypic distribution change over time.

The explanatory or numerical values of evolutionary models can be used to explain the evolutionary changes in the mimetic samples and target phenotypes and predict their phenotype distribution at different timepoints and environmental conditions. These predictions are used to test hypotheses and infer key factors in the evolutionary process. During evolution, genetic variance undergoes adaptive and dynamic changes according to fitness, thereby regulating the similarity between the source audio and the conditional audio in detail, so that the source audio can better learn the distribution of the conditional music. Under the approximate condition of weak selection, the final phenotype distribution tends to go toward the optimal under weak selection.

Ecologically driven coevolutionary dynamics enable us to use the weak selection approximation to derive a few general predictions regarding the evolution of phenotype mean and variance. Typically, the evolution of the mean phenotype reflects a balance between the selection pressures generated by direct stabilizing selection and those generated by interspecies interactions. For internal equilibria (i.e., equilibria with intermediate mean values), an approximate mean distance between two traits can be derived from the equations with the assumption of an asymmetric phenotype distribution. Then, at equilibrium, the mean phenotype will reach the physiologically optimal difference to interplay, which depends on the relative strength of the interaction in the mimetic sample or target. When mimetic evolution equilibrium is reached,

\[
\frac{\bar{\theta}_y}{\bar{\theta}_x} > \frac{|\bar{\theta}_x| - \bar{\theta}_x}{2L_x + 1|\bar{\theta}_x| + \bar{\theta}_x},
\]
(27)

\[
\hat{R} \equiv \frac{\bar{G}_x|\bar{\theta}_x + \bar{\theta}_y|}{\bar{G}_y|\bar{\theta}_y + \bar{\theta}_y|} < \frac{L_x}{2L_x - 1},
\]
(28)

where \(\hat{R}\) can be interpreted as the ratio of the potential evolutionary rates between the two audio signals. Through multiple rounds of statistical analysis of weak selection mimetic parameters, we found that setting \(\bar{\theta}_x = 0.002\), \(\bar{\theta}_x = -0.004\), \(\bar{\theta}_y = 0.004\), \(\bar{\theta}_y = 0.008\) achieves a balance between the diversity and fitting speed of evolutionary equilibrium. At equilibrium, the difference of the mean phenotype with respect to the physiologically optimal value is
By selecting a smaller coefficient, the rate of change in the frequency of trait genes in pseudo-populations is limited, which restricts the evolutionary rate of traits based on weak selection. This approach helps prevent the algorithm from producing unrealistic evolutionary paths by avoiding significant changes to traits that are only slightly favored by selection. This formula balances the mutual information between the mimetic sample and the target phenotype, thereby guiding the direction of conditional generation of the source audio. This can enable the source audio to learn the complex hierarchical structure of natural audio in a more natural and appropriate manner, while maintaining an appropriate balance between convergence and diversity. Thus, the generated natural sleep music achieves optimal weak selection mimetic differences concerning natural music, avoiding pitch mutations or unreasonable melodies, to obtain the best musical performance.

Furthermore, the flowchart of the novel WSM algorithm is given in Figure 4.

![Flowchart of the WSM algorithm](image-url)
3.2.5. Comparison with Evolutionary Algorithms

Evolutionary algorithms (EAs) [71] are optimization algorithms that simulate biological evolution to solve complex optimization problems. On the basis of Darwin’s theory of evolution, EAs progressively optimize solutions by simulating natural selection, inheritance, and mutation processes. To highlight the superiority of our weak selection mimicking algorithm (WSMA) in natural hypnosis music generation, we added genetic algorithm (GA) [72], distribution matching algorithm (DMA) [73], differential evolution algorithm (DEA) [74], cooperative coevolution algorithm (CCEA) [75], nondominated sorting genetic algorithm II (NSGA–II) [76,77], and weak selection mimicking algorithm (WSMA) in the model and compared their fidelity and diversity in natural hypnosis music generation [55,78]. We found that DMA, NSGA–II, and CCEA struggled to converge, resulting in missing details and elements in generated samples and lower fidelity. GA had intuitive and simple parameter settings and high computational efficiency, but it only imitated the audio part of the music, missing its musical properties. DEA emphasized matching part of the probability distributions and might be more suitable for specific music style transfer tasks, but only learned some musical attributes and ignored others. In contrast, WSMA could overcome these problems by simulating multi-objective distributions while ensuring music fidelity. Moreover, we compared the computational efficiency using the genetic algorithm (GA) as a baseline (five out of 10) for both fidelity and diversity. The corresponding visualization can be seen in Figure 5.

![Figure 5](image)

**Figure 5.** Fidelity (y-axis), diversity (x-axis), and computational efficiency (size of circles) of different algorithms.

As shown in Figure 5, the diversity and fidelity of natural hypnosis music generation are crucial for enhancing the listening experience, personalization, therapeutic applications, creative inspiration, and cultural representation. Our weak selection mimicking algorithm (WSMA) demonstrated superior performance in terms of both fidelity and
diversity, making it a promising approach for generating high-quality hypnotic music. With its ability to simulate multi-objective distributions while ensuring music fidelity, WSMA has the potential to revolutionize the field of natural hypnosis music generation. Its applications extend beyond entertainment and music production, with potential therapeutic benefits for individuals suffering from anxiety, stress, and sleep disorders.

Under the same crossover probability and mutation probability conditions, the WSMA is capable of avoiding local optima to a certain extent, thereby allowing for the evolution of weakly selected optimal solutions in multi-objective problems. On the other hand, other algorithms are prone to get trapped in local optima, resulting in slow convergence. This indicates that the weak selection principle employed by WSMA in solving multi-objective problems enhances the diversity of the initial state. The use of an adaptive genetic evolution probability allows for a balance between diversity and fidelity while adjusting the average phenotypic strategy enhances the algorithm’s global feature learning capabilities. Furthermore, the local feature learning strategy provides a directional evolution of features and improves precision in the search for optimal solutions in their proximity. To allow for a more intuitive comparison of the convergence abilities of WSMA and several other algorithms, Figure 6 illustrates the evolution curve of the average fitness values of nine music attributes.

![Figure 6](image.png)

According to Figure 6, it can be observed that WSMA has a relatively low fitness at the initial stage, which is advantageous for finding better attribute values. After multiple iterations, the fitness curve of WSMA is always higher than that of other algorithms. In the process of executing 100 iterations, WSMA is able to efficiently find approximate
solutions that meet the requirements, demonstrating a faster optimization speed. In contrast, other algorithms require more iterations to converge to the optimal solution, especially in the target attributes (e), (h), and (i), where the fitness convergence curves change slowly and may get trapped in local optima. WSMA adopts adaptive and multi-objective optimization strategies, which effectively avoid being trapped in local optima and continue searching for the global optimal solution. Therefore, using WSMA can improve the convergence speed and stability of the algorithm.

3.3. Overall Framework of the Weak Selection Mimetic Diffusion Probability Model (WSMDPM)

In this section, the proposed WSMDPM is elaborated with implementation details, whose overall framework is illustrated in Figure 7.

![Figure 7](image-url)

**Figure 7.** The network architecture of WSMDPM in modeling \( \epsilon_\theta: \mathbb{R}^k \times \mathbb{N} \to \mathbb{R}^k \).

The basic structure of WSMDPM comprises a residual stack consisting of \( N \) residual blocks, each having \( C \) residual channels, as illustrated in Figure 7. Firstly, the initial input \( x_T \) for the forward process of diffusion is obtained, following a Gaussian distribution. During the training process, the timestep \( t (t \in [T, \ldots, 1]) \) is conditioned upon, and music with different levels of noise is used as the input \( x_T \), which is a noisy audio signal. The weak selection approximation algorithm is employed to compute \( p_\theta(y_i|x_T, t) \) on \( x_T \) (where \( \theta \) represents the parameters of the model), and the gradient \( \nabla x_T \log p_\theta(y_i|x_T, t) \) is utilized to guide the diffusion sampling toward the target audio \( y_i \). Guided by the encoded information \( y_i \) in the target audio, noise is iteratively removed at each timestep to transform the current encoding \( x_i \) into \( x_{i-1} \), ultimately resulting in the generation of controllable natural hypnosis music \( x_n \). Specifically, the intermediate features of the encoded information \( x_i \) are enhanced through weak selection approximation, and transformations are performed on the target distribution \( y_i \) to achieve feature evolution. During evolution, when the evolution inertia \( s > 0.7 \), the distribution becomes sharper than
Mathematics 2023, 11, 3345

$p_0(y_t|x_t,t)$ because larger values are exponentially amplified, thereby learning global features. In other words, using a larger gradient scale, more attention is given to the style of the target, leading to higher fidelity (but less diversity) of the samples. When the evolution inertia extends to 0.4, the effect balances recall (diversity measure) to obtain higher precision and sample diversity. Ultimately, the optimal values are achieved at weak selection points.

Figure 8 illustrates the crucial role of the “Conditioner” module, which is also referred to as the condition module, in determining the output of our system. This module combines dilated convolution with conditional convolution by using two layers of transposed 2D convolution in the time and frequency domains. For each layer, the period span of upsampling is 16, and the 2D filter size is $[3, 32]$. After the upsampling operation, the Mel spectrogram is mapped to $2 \times C$ channels using a CONV $1 \times 1$ convolutional layer. A weak selection mimicking algorithm is added before the gate-tanh nonlinear activation function of each residual block. This algorithm computes the mutual adaptivity of the current input step $x_t$ and the prior $y_t$ of the Mel spectrogram and computes the effectiveness of planning transition strategies and the probability of controlling the feature transition to obtain a new update strategy $\frac{dy_t}{dt}$. If the evolutionary inertia $\frac{dy_t}{dt}$ is large, the learned conditional audio is a global feature. Conversely, the learned features are detail-oriented. By acquiring the variable $x_t$ that mutates each time the evolutionary equilibrium is attained, the conditional features can be adaptively enhanced to improve reconstruction accuracy.

![Figure 8. The structure of the conditional module.](image)

Furthermore, WSMDPM has an advantage in enlarging the receptive field of the output $x_0$; by iterating backward from $x_T$ to $x_0$, the size of the receptive field can be increased to $T \times r$, which makes WSMDPM suitable for conditional generation while reducing parameterization and minimizing information loss. This model adopts a non-autoregressive structure, which is different from the autoregressive structure of Wavenet. Therefore, it can avoid the problem of generating samples that are overly similar and
improve the diversity of generated samples. Directly inputting the Mel spectrogram into
the weak selection mimicking algorithm introduces additional high–frequency noise,
which results in a mismatch with the original signal $x_t$. To mitigate this issue, a condi-
tional module is proposed to constrain the high–frequency components of the target sam-
ple and accelerate the convergence speed of the algorithm while maintaining the gener-
ated music frequencies in a low–frequency range.

3.4. Conditional Module

A new conditional module is proposed to enhance the weak selection mimicking
generating ability of the model and improve the capacity of traditional diffusion models
in conditional audio generation. Specifically, a module with a residual structure is estab-
lished by designing a causal convolution layer module with a sufficiently large receptive
field. The unidirectional structure with strict time constraints in expanding causal con-
volution can effectively expand the receptive field of the convolution kernel in the time mod-
ule. This improvement enables the proposed conditional module to achieve (1) better res-
olution of the oversmoothed problem in Mel spectrogram generation caused by the mean
squared error or mean absolute error loss, (2) pursuit of fewer model parameters to avoid
overfitting, and (3) utilization of the residual structure to retain both shallow and deep
characteristics of the model. The structure of the proposed conditional module is pre-
SENTED in Figure 8.

The specific structures of the preprocessing and conditional modules are presented
in Figure 8. Di–Cond2d represents dilated convolution, BN represents batch normalization,
TransConv2d represents 2D transposed convolution, and the Leaky–ReLU function
is selected as the activation function. In the preprocessing procedure, the audio signal is
transformed into a spectrogram using short–time Fourier transform (STFT). Then, the Mel
spectrogram is upsampled using 2D transposed convolution to obtain waveforms of equal
length. After applying the Leaky–ReLU activation function, dilated convolution is used
to enlarge the receptive field, capture larger–scale image features, and avoid increasing
the original kernel size or introducing more weights. By applying convolution filters with
different dilation rates, the conditional block can capture features of multiple scales in a
low–cost and robust manner, as well as speed up the training process. However, the struc-
ture of the dilated convolution filter can cause feature information loss [43].

To maintain information and restore the complete spatial resolution of the network
output and facilitate the continuous updating of gradients in training, we used skip con-
nection feature maps and upsampled feature maps to sum for upsampled compensation
and bypass nonlinearity, thereby creating shortcuts. The resulting feature maps generated
from the conditional module are then used as inputs for the weak selection mimicking
algorithm.

4. Experiment and Discussion

4.1. Dataset and Implementation Details

4.1.1. Dataset

We systematically collected 100 instrumental songs on Spotify [79], which comprised
a playlist that is approximately 4 h long; this was used as the training dataset for the neural
network. In addition, conditioned audio inputs such as rain, wave sounds, wind sounds,
and water flow were used as inputs. The dataset was preprocessed by converting it into
Mel spectrograms.

4.1.2. Implementation Details

The audio was processed using a sample rate of 22,050 kHz and 80 Mel filters. A
short–time Fourier transform (STFT) with a window size of 1024, a hop size of 256, and an
FFT size of 1024 were applied. In the diffusion decoder, 30 residual blocks were used with
a convolutional channel size of 64 and a total diffusion time of 100. During training, an
Adam optimizer was used with a constant learning rate of 0.0002, and a dilation schedule of \([1, 2, \ldots, 512]\) was selected with diffusion factors \(\beta_t \in [1 \times 10^4, 0.02]\). The training batch was set to 4, and the step size was set to 800 steps.

4.2. Results

To evaluate the performance of our WSMDPM model in the context of hypnosis music composition, we generated 10 natural hypnosis music samples with conditioned inputs comprising natural sounds such as rain, waves, and birds chirping. Owing to the weak selection similarity diffusion process, the generated hypnosis music samples display a similarity to the original music with moderate variability, thereby showcasing greater creativity. When compared in terms of arousal level, the generated hypnosis music is characterized by low–frequency energy while also preserving the melodic features of the original natural sounds.

In order to evaluate the performance of our weak selection diffusion model in generating natural hypnosis music, we used the Spotify API [79] to analyze the characteristics of natural hypnosis music such as energy, pitch, and instrumentality. The features of the created natural hypnosis music are described below.

In terms of rhythm, natural sleep music is slower than regular music. These results indicate that natural sleep music has characteristics such as low energy, high instrumentality, low rhythm, and high pitch. It also includes the degree of variation in music rhythm, which conforms to the characteristics of meditation music. Additionally, the study found that natural sleep music covers different subgroups of natural audio features, thereby indicating that weak selection mimicry can evaluate the importance of different aspects of natural audio. The combination of selective representation of domain features and one’s own representation fully simulates the inherent properties and structure of natural audio, evolves into diverse mimicry, and supports adaptation to more complex music reasoning. As evident from Figure 9, the generated hypnosis music maintains high–level composition features such as tonality, chord sequences, and melody; simultaneously, meaningful low–level timbre texture features such as energy and loudness information are integrated into the ontology to achieve the optimal fitness phenotype. Therefore, weak selection mimicry is highly suitable for quantifying the fidelity of imitation on a continuous scale by identifying differences in perceptual structure and phenotype, thereby improving the quality of natural hypnosis music.

To induce synchronization of low–frequency neural activity or heart rate with the rhythmic structure of auditory stimuli, a weak selection mimicry algorithm was applied to adjust the distance between the generated music and the low–frequency distribution of natural audio. The rhythm of the generated music was adaptively slowed down to match the frequency range of \(\delta\) brainwave activity associated with deep sleep. Additionally, gentle natural audio was embedded into the music, and the pitch was adjusted using signal functions to align with the brainwave frequency of human slow–wave sleep [80]. The incorporation of slow variations in rhythm can enhance low–frequency activity in the brain, promoting sleep. Thus, this type of hypnosis music holds great promise for inducing sleep. Moreover, implicit suggestion plays a crucial role in the objective sleep intervention results of natural sleep music. It refers to the individual’s response to suggestions regarding perception, cognition, neural processes, and bodily functions [81]. This explains the strong individual differences in the effectiveness of hypnosis interventions that extend to phenomena such as the placebo effect [82]. The low suggestibility (nonverbal, highly implicit) of the natural sleep music we generated can interact significantly with slow–wave sleep (SWS) [83]. Its low suggestibility can increase participants’ percentage of SWS, indicating that music may have a stronger impact on the autonomic and central nervous systems during sleep.
Figure 9. Nine time–domain and frequency–domain characteristics of sleep music are presented in the form of smoothed density plots. The study found that, as the mean accumulates to higher values, the distributions of rhythm, loudness, energy, liveliness, and danceability features are skewed to the left. In contrast, the distributions of instrument and pitch features are skewed to the right, as their means accumulate to lower values.

Furthermore, to verify the performance of the proposed weak selection mimicry diffusion model by revealing the correlation among various audio features, similarity calculations were performed on nine identified secondary features. Figure 10 presents the difference heatmap between the generated hypnosis music features. To investigate whether or not there is a correlation between these variables and avoid multicollinearity, the autocorrelation matrix among the nine time–domain features in the music was visualized.
Figure 10. Correlation is represented by the intensity of color, with dark orange indicating positive correlation and dark green indicating negative correlation. It is evident that there is a strong positive correlation between liveliness and danceability features, which exists between liveliness and energy, as well as between rhythm and liveliness. On the other hand, there is a negative correlation between valence and tempo, as well as between valence and liveliness.

To further evaluate the effectiveness of the algorithm–generated natural sleep music, five tracks were selected for sleep tests. These tracks incorporate soothing sounds such as rain, waves, wind, bird songs, and insect chirping to create a calming atmosphere that is conducive to sleep. The Pittsburgh Sleep Quality Index (PSQI) [84] was used to evaluate the sleep quality (SSQ), sleep onset time (SL), sleep duration (SDu), and sleep efficiency (HSE) of 200 participants from different industries and age groups, including 113 male and 87 female individuals with sleep disorders. The total score was 12, with a lower score representing better sleep quality. Descriptive statistics were presented as mean ± standard deviation ($\bar{x} \pm s$). Compared with general hypnosis music, our method showed improvements in multiple aspects. The results are presented in Table 1.
According to the subjective evaluation of the participants in Table 1, both natural sleep music and general sleep music can improve sleep quality. However, natural hypnosis music has better sound quality and can fulfill human auditory needs. The sounds in natural hypnosis music are also closer to the sounds of the natural environment—such as bird song, flowing water, and rustling leaves—which can make people feel like they are amidst peaceful nature, have a relaxing and calming effect on them, thereby reducing inner anxiety and stress and helping people fall asleep faster. To adjust people’s breathing, we made the frequency characteristics of sleep music close to those of soothing music of 60–80 beats per minute, used the frequency characteristics of natural music to calculate feature vectors, and indicated weak selection approximation of feature vectors for generic music to generate rhythms that can approach the human heart rate (approximately 48–53 bpm). The frequency of 0.01–2 Hz in the music can increase slow-wave activity (SWA) and induce deep sleep, thereby significantly shortening the time it takes to fall asleep.

It can be observed from the table that, compared to generic sleep music, natural hypnosis music had significantly reduced scores in all aspects of PIQS and total scores. For example, after the natural hypnosis music intervention, sleep efficiency scores decreased by 0.25 and subjective sleep quality scores decreased by 0.28, thereby indicating that it can improve sleep efficiency and improve sleep conditions. This kind of natural sleep music has global therapeutic features and is closer to the physiological mechanism of humans, such as heart rate and breathing frequency. It is the most anticipated type of relaxation music. Specifically, it has low energy and danceability and high instrumentation and tonality, which represent the major common features of music used for sleep. The effectiveness of the weak selection approximation algorithm in generating music was demonstrated by comparing the Mel spectrograms of the source music, target sample, and approximated music, as depicted in Figure 11.

### Benchmarking Competing Models

Next, in order to compare the performance of WSMDPM for generating natural hypnosis music against existing music style transfer methods such as SleepGAN, CycleGAN, and TimbreTron [15,29,85], we ran an independent test where the WSMDPM is trained until convergence (around 800 iterations). Similar to the ablation study, source music was set as the content and target music was set as the style, and they were input into the style transfer model. By comparing the performance of different methods in terms of fidelity and quality, we could better evaluate their objective performance in generating natural hypnosis music. The fidelity and quality of music spectrograms were among the objective evaluation metrics that we used for this purpose. The results of this comparison are shown in Figure 11.

| Table 1. Comparison of PIQS scores between general sleep music and natural hypnosis music. |
| Scale Name | General Sleep Music | Natural Hypnosis Music |
| SSQ | 0.80 ± 0.43 | 0.52 ± 0.57 |
| SL | 0.85 ± 0.47 | 0.62 ± 0.46 |
| SDu | 1.19 ± 0.61 | 0.93 ± 0.37 |
| SE | 1.03 ± 0.67 | 0.78 ± 0.63 |
| PSQI | 3.87 ± 1.25 | 2.85 ± 1.36 |
Mathematics 2023, 11, 3345

![Spectrogram](source)
![Spectrogram](SleepGAN)
![Spectrogram](CycleGAN)
![Spectrogram](TimbreTron)

Figure 11. Upon comparing the Mel spectrograms of the source audio, simulated audio, and target audio, it is evident that the Mel spectrogram generated by WSMDPM exhibits lower-frequency features and finer details compared to other models based on music style transfer.

The timescale is represented on the X-axis and frequency is represented on the Y-axis. According to the findings presented in Figure 11, CycleGAN is an adversarial generative network (GAN) method used for style transfer. In the context of music style transfer, it is possible that the generated music contains missing spectral frames due to the inability of the model to capture all of the details and completeness in the spectrogram. Similarly, SleepGAN is a style transfer method targeted toward generating hypnotic music. This may be caused by the model’s incapability to accurately learn all the spectral frames from the input music, which results in missing frames in the generated music. On another note, TimbreTron is a WaveNet-based generative model that fails to maintain clarity and accuracy when performing timbre transfer, resulting in noise components being introduced into the generated music, which undermines its musical value.

However, the spectrograms of the naturally generated hypnotic music by WSMDPM do not exhibit any missing frames or noise, thereby enhancing the quality and authenticity of the generated music. Moreover, our natural hypnotic music generates a Mel spectrogram with a more prominent low-frequency range, indicating that the low-frequency component plays a dominant role and achieving spectral attenuation to align the generated music more closely to human physiological rhythms. The overall low frequency may result in a longer sleep induction time and a greater likelihood of slow-wave sleep. Furthermore, the generated natural hypnosis music can reduce spectral distortion and achieve more accurate pitch prediction, thereby increasing the novelty and interest of the music. This indicates that the weak selection approximation music generation algorithm selectively and reasonably evolved the mean and variance of music features, thereby contributing to creating a better music experience and causing the generated music sound closer to real music.

5. Conclusions

In this paper, a new weak selection mimicking algorithm was proposed and successfully applied to the task of generating natural sleep music. The algorithm adaptively
controls the model’s ability to model global and local features by calculating the conditional evolutionary momentum in the conditional diffusion probability model, thereby striking a balance between the diversity and quality of generated natural sleep pressure. To accelerate the convergence speed of the weak selection similarity algorithm and meet the requirements of sleep music features, a novel conditional module is introduced before the algorithm, which can generate local embedding of conditional audio in a low-cost and robust manner by utilizing features at multiple scales and eliminating high-frequency components in them. It is worth noting that the PSQI was used to compare the effects of natural sleep music and general hypnosis music. The experiment revealed that our model effectively learns the low-frequency melody of natural audio and can approximate the rhythm of music with human physiological rhythms (such as blood pressure and respiratory rate), which is closer to the breathing and heart rate frequencies of humans, thereby reducing arousal reactions, meeting people’s preferences for natural environmental sounds, helping people relax their minds, and promoting deep sleep. According to subjective and objective evaluations, our music is more in line with human sleep physiology.

6. Future Work

In future research, combining the brainwave patterns of different participants to generate hypnotic music that aligns with individual physiological indicators is a promising direction. This approach would allow each individual to experience the most suitable relaxation and sleep outcomes. Furthermore, incorporating participants' preferred hypnotic music preferences can generate more personalized hypnotic music, enhancing diversity and providing a wider range of choices for individuals. However, the weakly selected mimic algorithm's computational process may be somewhat slow due to its complexity. Therefore, exploring methods for parallel processing of audio data in the future can enhance computational efficiency. These suggestions contribute to the further development of natural hypnotic music and offer new directions and possibilities for future research.

Author Contributions: Conceptualization, W.H. and F.Z.; methodology, W.H. and F.Z.; software, W.H. and F.Z.; validation, W.H. and F.Z.; formal analysis, F.Z. and W.H.; investigation, F.Z.; resources, F.Z. and W.H.; data curation, W.H. and F.Z.; writing—original draft preparation, F.Z.; writing—review and editing, W.H.; visualization, F.Z.; supervision, W.H.; project administration, W.H.; funding acquisition, W.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Guangzhou Youth Science and Technology Education Project (No. KP2023245).

Data Availability Statement: Not applicable.

Acknowledgments: The authors want to thank Guangzhou University for its support.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Overview of Spotify API music functionality.

<table>
<thead>
<tr>
<th>Audio Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loudness</td>
<td>A value indicating the overall loudness of a track, ranging between –60 and 0 dB. Spotify does not specify a dB scale, but it is assumed that it is measured in LUFS (loudness units relative to full scale).</td>
</tr>
<tr>
<td>Energy</td>
<td>A value indicating a perceptual measure of intensity and activity, ranging between 0 and 1.</td>
</tr>
<tr>
<td>Acousticness</td>
<td>A value indicating how likely it is that a track is acoustic, which means that it is performed on non-amplified instruments, ranging between 0 and 1.</td>
</tr>
<tr>
<td>Instrumentalness</td>
<td>A value indicating how likely it is that a track contains no vocals, ranging between 0 and 1 with values above 0.5 likely to be instrumental tracks.</td>
</tr>
</tbody>
</table>
Danceability
A value indicating how suitable a track is for dancing, ranging between 0 and 1, with higher values indicating increased danceability.

Valence
A value indicating positively valenced a track is (from a Western perspective), ranging between 0 and 1.

Tempo
A value indicating the speed or pace of track, as estimated by the average beat duration, given in beats per minute (BPM).

Liveness
A value indicating how likely a track was performed live, e.g., by identifying the sound of an audience in a recording.

Speechiness
A value indicating the presence of spoken words in a track.

References


**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.