人工知能学会第二種研究会資料
AIチャレンジ研究会
SIG-Challenge-066-09

A Comparison of generative models for constructing evolutionary models of bird vocalization

Hao Zhao[1*] Reiji Suzuki[1] Kazuhiro Nakadai[2] Takaya Arita[1]

[1] Nagoya University
[2] Institute of Science Tokyo

Abstract: We investigated and compared two agent-based evolutionary models for sexual selection of Blue-and-white Flycatcher songs, based on Variational Autoencoder (VAE) and Text-To-Audio (TTA) models. In the TTA evolutionary model, we extended the numerical genotypes from the previously proposed VAE-based model to textual genotypes for generating audio using Stable Audio Open 1.0, and applied mutations using a large language model (Gemma-2). The preliminary results of evolution experiments indicate that while the VAE evolutionary model can drive differentiation in genotypes and phenotypes under sexual selection, it tends to produce homogenized evolved songs due to inherent selection pressure favoring well-reconstructed and clear vocalizations near the latent space origin. In contrast, the Text-To-Audio evolutionary model promotes greater genotype diversity and fosters the creation of novel genes, while preserving the core characteristics of the original Blue-and-white Flycatcher songs. This allows the model to balance genotype and phenotype differentiation, demonstrating its potential to effectively capture the evolutionary complexity of natural vocalizations.

1 INTRODUCTION

Agent-based modeling has contributed significantly to our understanding of the evolutionary dynamics of social behaviors, particularly the emergence of communication through various signaling mechanisms. These models demonstrate how simple interaction rules among agents and population-level evolutionary processes can yield complex evolutionary phenomena. However, the nature and complexity of signals in these models often differ from those observed in real-world communication systems.

Deep learning techniques are contributing to computational bioacoustics [1], and generative models have recently been used in the study of animal communication and ecoacoustics. Variational Autoencoders (VAEs) [2] have demonstrated effectiveness in modeling latent variables in both experimental and real-world acoustic data [3] and have recently shown promise in synthesizing bird songs [4]. Gibbons et al. explore generative AI models, such as Auxiliary Classifier Generative Adversarial Networks (ACGAN) and Denoising Diffusion Probabilistic Models (DDPMs), to enhance bioacoustic species classification, particularly in challenging, noisy environments [5].

Recently, another emerging generative audio model, Text-To-Audio (TTA), such as models like AudioGen [6] and Stable Audio [7], has recently gained attention for its ability to synthesize general audio based on textual descriptions. While it has progressed in music generation and human speech creation, its application to animal sound generation remains limited.

Figure 1: Recording field.

Also, with the rapid development of large language models (LLMs), their generative capabilities can be used in the process of genetic mutation within evolutionary models. Fernando et al. proposed a framework that utilizes prompts to describe mutation methods, effectively enhancing genetic evolution [8].

Our objective is to extend agent-based evolutionary models by harnessing the rich and realistic generative capabilities of generative models and to explore potential interactions with real ecological systems. As a preliminary approach, Suzuki et al. constructed a novel agent-based evolutionary model for animal vocalizations utilizing generative models, exemplified through case studies on several species (e.g., Japanese Bush-warbler (*Horornis diphone*) [9], Spotted Towhee (*Pipilo maculatus*)[10]). They focused

*Nagoya University
Furo-cho, Chikusa-ku, Nagoya 464-8601

E-mail:zhao.hao.y0@s.mail.nagoya-u.ac.jp

on a sexual selection process inspired by Higashi et al. [11] where male song and female preference genotypes were represented as vectors within the 2D latent space of a Variational Autoencoder (VAE) trained on the focal species. Spectrogram images generated from these vectors were interpreted as vocalizations and song preferences. Females probabilistically choose males based on the similarity between the spectrogram images of male songs and their own preference spectrogram images. The results indicated that clear and moderately complex vocalizations were preferentially selected during the evolutionary process and sometimes exhibited segregation of the population, which may lead to sympatric speciation. However, their evolutionary model still faces several limitations. First, the evolutionary process is strongly constrained by the limitations of the latent space, leading to rapid population convergence. Second, the diversity in generated songs (phenotypes) is restricted to interpolations within the training data space.

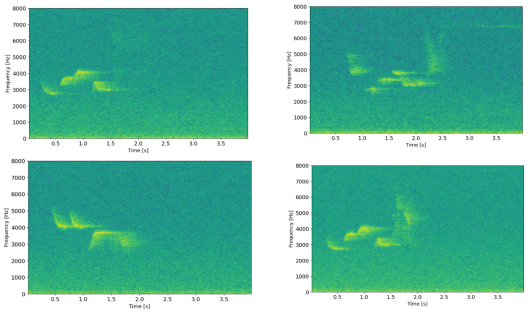

Figure 2: Original sound clips of Blue-and-white Flycatcher.

This study employs a Text-To-Audio (TTA) generation model to investigate the evolutionary dynamics of vocal signals and preferences in sexual selection. It incorporates a more flexible genetic space based on linguistic expressions and diverse vocal representations. As an initial step, we replace the phenotype mapping component in Suzuki and Arita's sexual selection model, which originally used a VAE with numerically encoded genes, with a TTA generation system that employs natural language-based genetic representations.

We fine-tune the Text-To-Audio model, Stable Audio Open 1.0 (Stability AI) [7], using a dataset of Blue-and-white Flycatcher (BWFL) songs. The trained model is expected to generate songs that share characteristics with, but are not identical to, those of the focal species, based on various textual descriptions of the species' songs. These descriptions serve as genes encoding both male songs and female preferences. Additionally, we utilize a Large Language Model (LLM), Gemma-2 (Google) [12], to express genetic mutations in textual form. We first demonstrate how these models generate songs resembling BWFL songs. Then, through preliminary evolutionary experiments using these models, we examine the effects of different generative models on evolution within individual-based evolutionary models.

2 METHODS

2.1 Field recording and dataset

Field recordings of the Blue-and-white Flycatcher were conducted in the Inabu Field, an experimental forest of the Field Science Center, Graduate School of Bioagricultural Sciences, Nagoya University, in central Japan (Fig. 1). The recordings were conducted on June 5th, 2024, using a 16-channel microphone array system (Chirpy type-S; System in Frontier Inc.) in 16-bit, 16 kHz format. We manually selected 100 sound clips of the songs from the recording (11:00 AM to 11:20 AM on June 5, 2024) as the training samples (Fig. 2). The duration of each clip was 4 seconds.

2.2 Generative audio models

We constructed two generative audio models using the dataset above. The construction process is outlined as follows.

2.2.1 Training Variational Autoencoder model

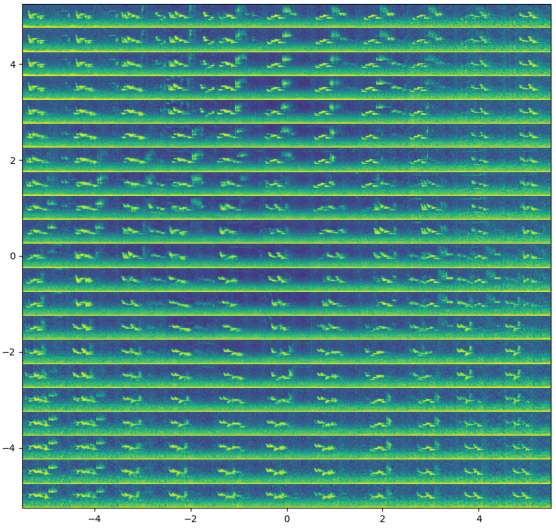

Figure 3: Song distribution within the latent space of Blue-and-white Flycatcher songs, with feature vectors represented by spectrograms. Songs near the origin in the center are most similar to natural songs.

The first model, built from the recorded songs, employs a convolutional variational Autoencoder (VAE),

as described by Suzuki et al. [10]. This model features an encoder with eight convolutional layers and three fully connected layers that compress the information into two dimensions, along with a decoder that mirrors the architecture of the encoder [13].

The 100 4-sec sound clips were converted into 496 x 128-pixel grayscale sound spectrogram images and used as the training dataset. A representative spectrogram was generated from the corresponding coordinate position in the 2-dimensional latent space and mapped onto the same coordinate system (Fig. 3).

2.2.2 Fine-tuning Text-To-Audio model

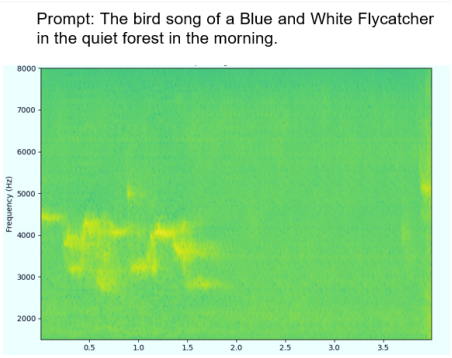

Figure 4: An example sound generated using the fine-tuned Stable Audio Open 1.0 model, with the prompt "The bird song of a Blue and White Flycatcher in the quiet forest in the morning.".

The second model, introduced in this study, involves fine-tuning the pre-trained Stable Audio Open 1.0 (Stability AI). This Text-to-Audio generative model utilizes transformer diffusion techniques, allowing it to produce artificial sounds based on text prompts [7].

We used the audio files of the above 100 sound clips. We paired these files with a text description (prompt), "The bird song of a Blue and White Flycatcher in the quiet forest in the morning," for training across 30 iterations. Fig. 4 illustrates an example of an artificial bird song generated by this fine-tuned model with this text prompt. While preserving the original timbre, the generated sounds seem to incorporate features from multiple original song samples.

Fig. 5 also shows several examples of the song spectrograms generated by the models before and after fine-tuning using different additional prompts with various descriptions of songs, such as “lonely, and quiet". The generated sounds with the original model appear to be vocalizations of some species, but they differ from those of BWFL. After fine-tuning, the generated songs have diverse and unique acoustic structures while retaining the properties of BWFL songs. In addition, while it is not clear, there seems to be some similarity in the sound structures (e.g., temporal changes) between the sounds produced with the same prompt. This indicates that the additional descriptions in each prompt, at least in part, are reflected in the acoustic properties of the generated songs, which implies that different prompts can bring about diverse BWFL-like songs.

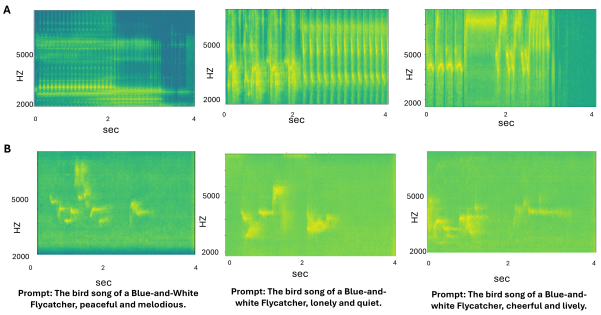

Figure 5: The generated songs from the original and fine-tuned TTA models. A) Before training with the same prompts.; B) After training.

2.3 Evolutionary model of bird songs using the generative audio models

First, we describe the original evolutionary model by Suzuki et al. [10], which uses a VAE. Then, we present the extended version incorporating a Text-To-Audio model and a large language model (Fig.6).

2.3.1 Evolutionary model based on VAE

In this evolutionary model (Fig.6 (left)), each individual has two real-valued genes. We assume the two populations, each composed of N males and N females, respectively. Each gene represents a 2D vector (or position) in the latent space, described by a pair of (x, y) coordinates (Fig.3). One gene is used to generate a song spectrogram vocalized by a male, and the other is used to generate a female preference spectrogram using the VAE with (x, y) as the latent vector. In the initial population, both genes' x and y coordinates are generated randomly within the range $[-W, W]$. Females select one male from all males with a probability proportional to $\exp(-\beta \cdot x)$, where x is the average difference in pixel values between the male's song spectrogram and the female's preference spectrogram, and β is a coefficient. Therefore, females stochastically select the male whose song is closest to their preference spectrogram. One male and one female offspring are produced from the parental genes, with recombination and mutation effects included. Recombination is modeled as BLX-α crossover [14] with a probability p_c, a crossover method designed to produce offspring genes by combining the characteristics of the parents' real-valued genes within a defined range. Mutation is modeled as a normal random value with a mean of 0 and a standard deviation σ, occurring in each gene with a probability p_m. The trials are conducted over T generations.

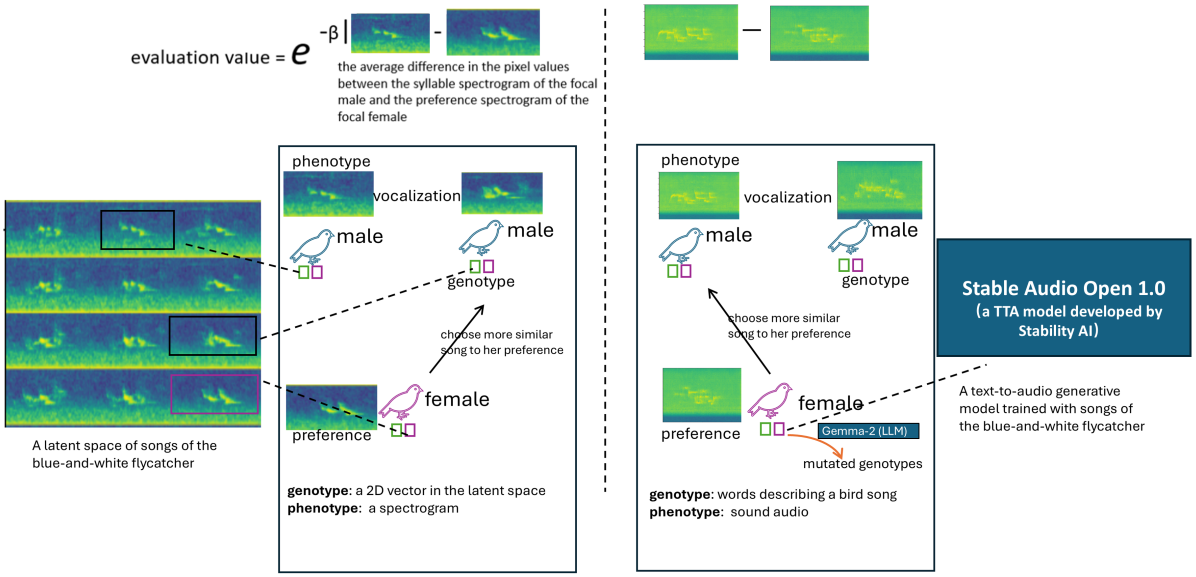

Figure 6: Evolution models based on VAE and Text-To-Audio model.

2.3.2 Evolutionary model based on Text-To-Audio model

Unlike the VAE-based evolutionary model, each individual consists of two three-word text genes describing some properties of songs (e.g. "sweet, bright, complex"). One is a gene used to generate a song vocalized by a male, and the other is a gene used to generate a female preference song (Fig.6 (right)). The gene expression involves utilizing the vocalizing gene or preference gene to construct a prompt for generating a song of BWFL. The prompt is formatted as:

> *The bird song of a Blue-and-white Flycatcher, which is {word1}, {word2}, and {word3}.*

This prompt is used to generate a corresponding song using the fine-tuned model of Stable Audio Open 1.0, as mentioned above. This prompt can help the generated songs closely resemble the original songs of the Blue-and-white Flycatcher while being different from them. It should be noted that our goal is not to create songs that directly match the described properties, but rather to use the creative ability of the TTA model to generate novel BWLF-like songs guided by these descriptions.

A mutation is performed on each text gene of parents using a large language model named Gemma-2-9b-it, which is a 9-billion-parameter instruction-tuned model developed by Google and available from Hugging Face [1]. The mutation was guided by a defined prompt with a probability p_m. The used prompt is as follows:

> Please partially modify the following three words to describe a bird song. The output must consist of exactly three words formatted as: `word1, word2, word3`. Do not include explanations, introductions, or additional symbols. The original description is: `{text gene}`.

Recombination will exchange partial words between the two vocalizing genes or the two preference genes of the parents with a probability p_c. Each gene in the initial population was randomly generated by the LLM as three words describing a birdsong.

3 Results

3.1 Evolutionary experiment based on VAE model

For the VAE evolutionary model, we used the parameter settings as follows: $N = 20$, $W = 5.0$, $\beta = 0.3$, $\alpha = 1.1$, $T = 80$, $p_c = 0.5$, $p_m = 0.15$, and $\sigma = 0.2$. The analysis of several trials showed that the male vocalizing genes and female preference genes tended to cluster one (Fig. 7 C) or two groups (Fig. 7 B) from the initial population in each trial. Both song and preference genes converged to similar positions. The distribution of genes varied significantly across trials, so additional experiments were conducted to identify overall trends. Fig. 8 (A) shows the frequency distribution of the vocalization genes of the last generation of males in 100 trials, using a kernel density estimation (KDE). Fig. 8 (B) shows a measurement of the Acoustic Complexity Index (ACI) [15] for the spectrograms in the latent space. The Acoustic Complexity Index (ACI) measures temporal changes in sound intensity across frequency bins. High ACI values are complex or noisy temporal variations in sound intensity within

[1] https://huggingface.co/google/gemma-2-2b-it

frequency bins, while low ACI values represent more consistent sound intensity patterns over time.

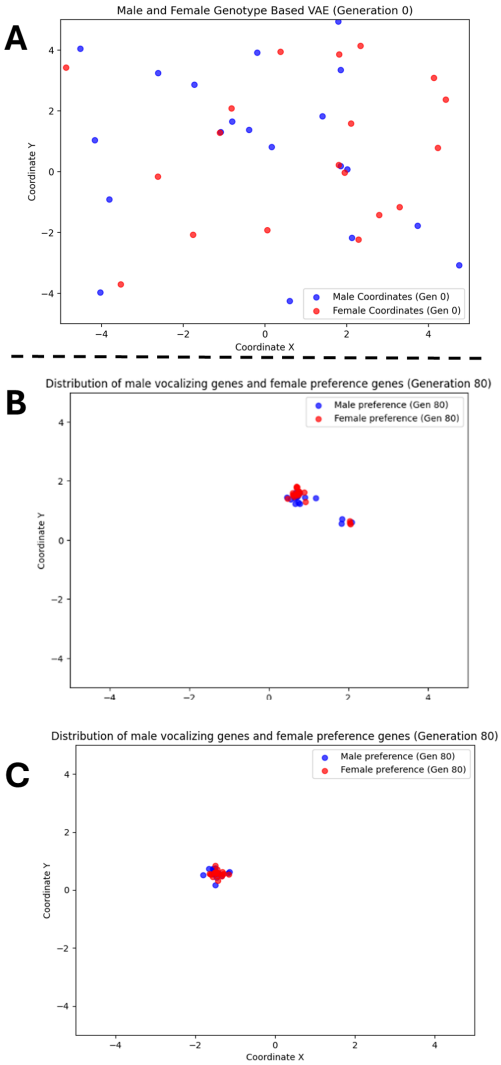

Figure 7: Distribution of male vocalization genes (blue) and female preference genes (red) in the VAE evolutionary model: A) initial population; B) and C) final population from different trials.

The frequency distribution of vocalization genes (Fig. 8 A) shows that the selected songs are generally distributed over a wide area around the origin. Compared to the ACI distribution (Fig. 8 B), the evolved songs tend to be distributed in a range with less noise, and the vocalizations are generated relatively clearly or well reconstructed, indicating a relatively lower ACI. Suzuki et al. tested a comparative approach using the same model on Spotted Towhee songs [10], where mating was based on the distance between gene coordinates instead of spectrogram differences. This led to rapid convergence to an unimodal distribution centered at the origin in repeated experiments.

This suggests that in the VAE evolutionary model, the evolution of songs in populations faces an inherent selection pressure. This pressure may arise because songs near the origin of the latent space typically exhibit simple and stable characteristics corresponding to the average features of the training songs. As a result, the songs of populations tend to evolve within the central region of the latent space. However, the influence of complexity and uniqueness of generated vocalizations and preferences on the evolutionary process has, to some extent, mitigated the homogenization of population genes and contributed to increasing the diversity of evolved songs.

3.2 Evolutionary experiment based on Text-To-Audio model

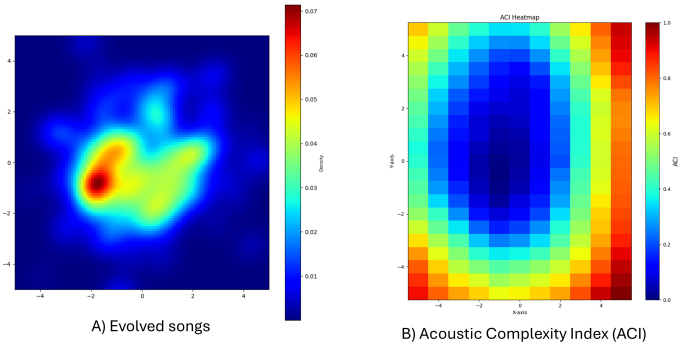

Figure 8: A) Frequency distribution of male vocalization genes in the final generation across 100 trials; B) Distribution of the Acoustic Complexity Index (ACI) of spectrograms in the latent space (Fig. 3). In both figures, red indicates higher values, and blue indicates lower values.

For the Text-To-Audio evolutionary model, we used the parameter settings as follows: $N = 20$, $\beta = 0.3$, $T = 80$, $p_c = 0.5$, $p_m = 0.15$. We compared the distribution of male vocalization genes and female preference genes between the initial and final populations in a pilot experimental trial. Both genes, represented as text, were transformed into 2D vectors using the SentenceTransformer package in Python and Uniform Manifold Approximation and Projection (UMAP). Fig. 9 illustrates the distribution of male vocalization genes and female preference genes in the 2D latent space for the initial and final population.

We observed that the distribution of male vocalization and female preference genes in the latent space shifted during the evolutionary process, deviating from their initial positions and forming clusters (Fig. 9 (right)). Both vocalization and preference genes were organized into several discrete groups. For example, both male song and female preference genes with "trilling" formed distinct clusters on the top right area. This suggests that the evolutionary process drives changes in male vocalization and female prefer-

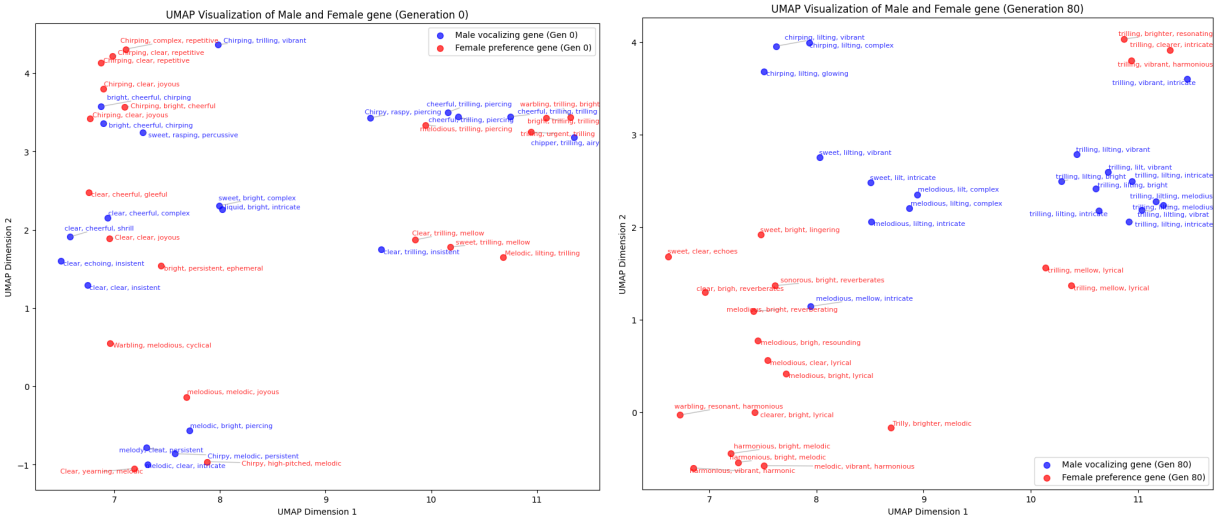

Figure 9: Distribution of male vocalization genes (blue) and female preference genes (red) in the initial (left) and final (right) populations within the latent space, based on the Text-To-Audio evolutionary model. Note that the two figures show the different distributions on the same latent space.

ence genes, leading to clustering tendencies under the rules of sexual selection [11] due to some kind of runaway process. Meanwhile, the mutations based on the LLM contributed to the creation of novel genes (e.g., the preference gene “trilling, brighter, resonating" and the vocalization gene “chirping, lilting, glowing")(Fig. 9 (right)), thereby enhancing the genetic diversity of the population. Consequently, the gene clusters are distributed across distinct regions of the latent space compared with the initial population.

However, the clusters of vocalization genes and the clusters of preference genes are not closely located in the latent space, implying that complex relationships among genotypes and phenotypes might have strongly affected the evolution process. We used the wav2vec 2.0 model [16] to extract the evolved vocalization and preference song embeddings in the initial and final populations and visualize them in the 2D latent space (Fig. 10).

The latent space distribution of male vocalization and female preference songs maintained a consistent structure across generations, likely reflecting inherent characteristics of BWFL songs. Both phenotypes gradually mixed together in the latent space through sexual selection, indicating increased affinity between vocalization and preference songs. However, their corresponding genotypes showed no clear spatial proximity among neighbors. This discrepancy suggests complex genotype-phenotype relationships that make the selection process non-trivial. This is because distinct genes could produce structurally similar songs in the latent space, as demonstrated by the proximity of songs generated from different descriptive sequences (e.g., “chirping, lifting, complex" for vocalization and “harmonious, vibrant, harmonic" for preference) (Fig. 10 bottom).

These results indicate that in the Text-To-Audio evolutionary model, the influence of LLM promotes greater diversity in the population’s genotypes, which are further differentiated during the evolutionary process. However, the evolved songs still face evolutionary pressure, as gene expression guided by specific prompts and the structure of the original songs ensures that the evolved songs retain the characteristics of the original Blue-and-white Flycatcher songs. Consequently, while different vocalization and preference genes produce songs with distinct properties, they also maintain the overall features of the original songs. This effectively preserves the core structure of the evolved songs within the population. The model might balance the differentiation of population genotypes (vocalization genes and preference genes) with the differentiation of phenotypes (evolved songs).

Therefore, different vocalization genes and preference genes may produce similar song structures, limiting the overall variation in the evolved songs.

3.3 Comparison of the evolutionary dynamics of genes between VAE and TTA-based models

To quantitatively compare the evolutionary dynamics of male vocalizing genes and female preference genes in both evolutionary models, we focused on the generational changes in the degree of clustering in both experiments. In both cases, we used the average cosine similarity of possible pairs among male vocalizing genes or female preference genes in the 2D space discussed above. The cosine similarity of these genes was calculated every five generations.

Fig. 11 shows the comparison of average cosine similarity across generations between the Text-To-Audio (left) and VAE evolutionary (right) models. In the

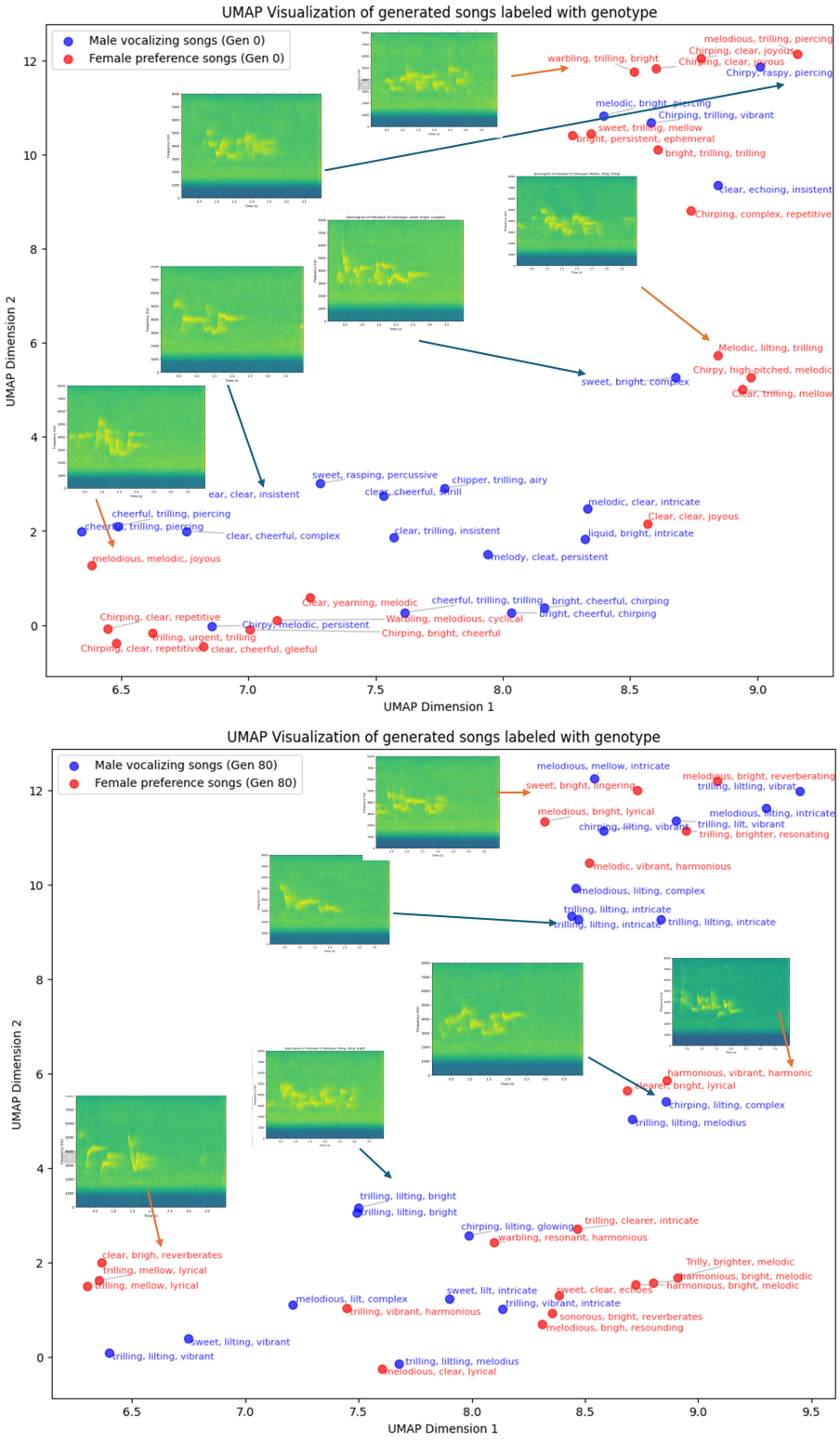

Figure 10: Latent space of evolved songs in initial (top) and final (bottom) population based on Text-To-Audio evolutionary model. Note that the two figures show the different distributions on the same latent space.

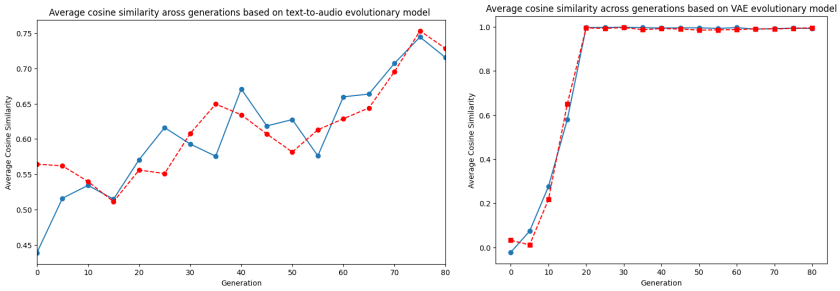

Figure 11: The average similarity of female preference genes (red) and male vocalizing genes(blue) across the evolutionary process. (left) Based on the Text-To-Audio evolutionary model; (right) Based on the VAE evolutionary model.

VAE model (right), the population rapidly converged, reaching a similarity of 1.0, with their genes becoming nearly identical. In contrast, the text-to-audio model (left) showed a more gradual increase in similarity over generations, suggesting maintained differences in the population's genes. This suggests that the large language model can enhance genetic diversity within the population during the evolutionary process, partially mitigating the converging trend of genes in the process of sexual selection, which still needs investigation.

4 Conclusion

In this study, we compared two evolutionary models of bird songs based on two generative audio models: the VAE and the Text-To-Audio models. The VAE evolutionary model, though capable of driving differentiation in genotypes and phenotypes under the influence of the evolutionary framework, tends to produce homogenized evolved songs due to the inherent selection pressure on well-reconstructed and clear vocalizations near the origin of the latent space. However, the Text-To-Audio model, along with the LLM-based gene mutations, promotes greater diversity of genotypes and fosters the creation of novel genes. This is due to more distinct song properties while maintaining the core characteristics of the original Blue-and-white Flycatcher songs. Therefore, the Text-To-Audio model balances the differentiation of genotypes and phenotypes, showcasing its potential to better simulate the evolutionary complexity of natural vocalizations.

While further experiments and analyses are necessary, these results highlight the advantages of integrating natural language models into evolutionary frameworks for studying and generating complex animal behaviors, specifically in realizing the evolutionary emergence of novel yet realistic signals in ecological contexts.

ACKNOWLEDGMENTS

This study is supported in part by JSPS KAKENHI 24K15103, JST SPRING JPMJSP2125, and Google Gemma 2 Academic Program.

References

[1] Dan Stowell. Computational bioacoustics with deep learning: a review and roadmap. *PeerJ*, 10:e13152, 2022.

[2] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[3] Aman Singh and Tokunbo Ogunfunmi. An overview of variational autoencoders for source separation, finance, and bio-signal applications. *Entropy*, 24(1):55, 2021.

[4] Axel-Christian Guei, Sylvain Christin, Nicolas Lecomte, and Éric Hervet. Ecogen: Bird sounds generation using deep learning. *Methods in Ecology and Evolution*, 15(1):69–79, 2024.

[5] Anthony Gibbons, Emma King, Ian Donohue, and Andrew Parnell. Generative ai-based data augmentation for improved bioacoustic classification in noisy environments. *arXiv preprint arXiv:2412.01530*, 2024.

[6] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022.

[7] Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. *arXiv preprint arXiv:2407.14358*, 2024.

[8] Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*, 2023.

[9] Reiji Suzuki, Shinji Sumitani, Chihiro Ikeda, and Takaya Arita. A modeling and experimental framework for understanding evolutionary and ecological roles of acoustic behavior using a generative model. Proceedings of ALIFE 2022: The 2022 Conference on Artificial Life (ALIFE2022), Paper No: isal_a_00542, 58 (3 pages), July 2022.

[10] Reiji Suzuki, Zachary Harlow, Kazuhiro Nakadai, and Takaya Arita. Toward integrating evolutionary models and field experiments on avian vocalization using trait representations based on generative models. In *Proceedings of 4th International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots*, pages 69–73, 2024.

[11] Masahiko Higashi, Gaku Takimoto, and Norio Yamamura. Sympatric speciation by sexual selection. *Nature*, 402(6761):523–526, 1999.

[12] Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[13] Tim Sainburg, Marvin Thielk, and Timothy Q Gentner. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS computational biology*, 16(10):e1008228, 2020.

[14] Larry J Eshelman and J David Schaffer. Real-coded genetic algorithms and interval-schemata. In *Foundations of genetic algorithms*, volume 2, pages 187–202. Elsevier, 1993.

[15] Nadia Pieretti, Almo Farina, and Davide Morri. A new methodology to infer the singing activity of an avian community: The acoustic complexity index (aci). *Ecological indicators*, 11(3):868–873, 2011.

[16] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.