EXPLORING PANSORI GENERATION WITH ACE-STEP

Seola Cho[1], **Minjun Kim**[2], **Dasaem Jeong**[1]
[1]Department of Art & Technology, Sogang University
[2]Department of Electrical and Electronic Engineering, Yonsei University
{seola.cho, dasaemj}@sogang.ac.kr, oxymoron@yonsei.ac.kr

ABSTRACT

Pansori is a traditional Korean narrative musical form that combines sung passages with spoken narration (*aniri*) under explicit rhythmic cycles (*jangdan*). While recent text-to-music models achieve strong results on mainstream genres, their behavior on underrepresented traditions remains largely unexplored. We adapt ACE-Step to pansori via LoRA-based fine-tuning and compare its adaptation to Korean pop as a mainstream baseline. We evaluate our results through quantitative analysis with Fréchet Audio Distance (FAD) and qualitative assessment by an expert pansori performer. Our study highlights challenges and partial successes in adapting to pansori. Future work will extend text-to-music models toward cultural diversity and traditional music.

1. INTRODUCTION

Recent advances in text-to-music models [1–3] from commercial services demonstrated that these models are capable of generating plausible music with proper styles in audio, including vocal with lyric condition. Following these services, open source models that can generate whole song audio have been introduced [4–7]. While these models show strong performance in generating popular music genres, their usability for generating traditional music of various countries is still under-explored [8, 9].

In this work, we investigate the adaptation of ACE-Step [7] to *pansori*. Pansori is a traditional Korean narrative musical form performed by a solo singer (*sorikkun*) [10]. It combines singing (*sori*) with spoken narration (*aniri*) under explicit rhythmic cycles (*jangdan*). The art form has its origins in shamanistic music of the *Namdo* region and evolved through the creativity of many master performers who contributed melodic variations known as *deoneum*. The transmission of pansori remains primarily oral, emphasizing the individuality of singers who interpret and innovate within this living tradition.

Pansori is actively being revitalized today through creative reinterpretations, new compositions, and audience-friendly performances. Particularly noteworthy are the sorikkun Jaram Lee's adaptations of Western literature (e.g., Ernest Hemingway's *The Old Man and the Sea*, Victor Hugo's *Les Misérables*). In such cases, there is significant potential for AI generation to assist in creating pansori music for new, non-traditional texts, contributing to the development and coexistence of traditional music.

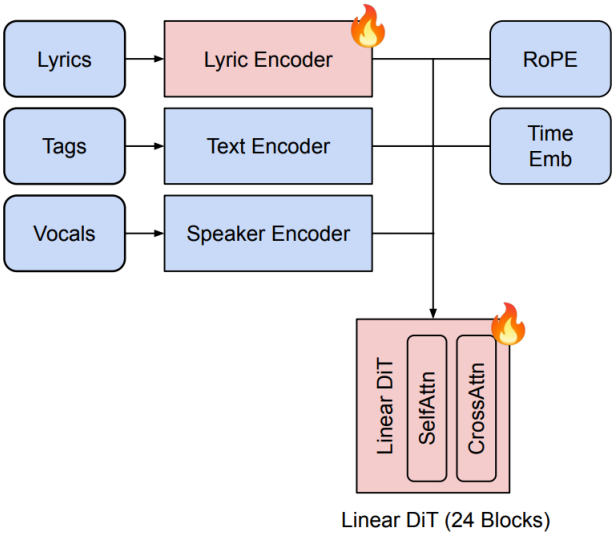

Figure 1. ACE-Step Overview

Given the complex melodic, rhythmic, and vocal heterogeneity intrinsic to pansori, including its oral tradition and lack of comprehensive symbolic notation, audio-based models are particularly suitable for pansori generation. Unlike symbolic music representation, audio models can grasp the subtle vocal timbres, ornamentation, and the interplay between sung and spoken parts that define pansori's authenticity.

This study explores the capabilities and limitations of adapting ACE-Step to pansori, and discusses their future directions for text-to-music models applied to traditional genres across cultures.

2. METHOD

2.1 Dataset

We used 7 hours of pansori audio performed by Jueun Lee, a lead artist of Folk Music Group of National Gugak Center, with permission from the performer. In this

Late-Breaking / Demo Session Extended Abstract, ISMIR 2025 Conference

work, we define *segment markers* as symbolic labels that include *jangdan*, rhythmic patterns performed on a drum when singing, and *aniri*, spoken narrative passages rather than sung parts. These markers were enclosed in square brackets in the lyrics to provide explicit cues for the model. Audio was segmented into clips along lyric lines. We ensured that each *segment marker* was preserved and followed ACE-Step's recommended maximum audio length of 4 minutes per clip. For comparison, we included 2 hours of Korean pop songs performed by Kwang-seok Kim.

2.2 LoRA Fine-tuning

The model was trained for 23,000 steps using LoRA [11] with the default configuration ($r = 256$ and $\alpha = 32$). As illustrated in Figure 1, LoRA is applied to the lyric encoder and the main transformer.

3. RESULTS

3.1 Analysis between Pansori and Korean pop

LoRA weights	Training steps				
	5k	**10k**	**15k**	**20k**	**23k**
1	0.929	0.771	0.821	0.812	0.643
2	0.601	0.481	0.430	0.544	0.423
3	0.655	0.626	0.596	0.765	0.648

Table 1. FAD↓ scores for pansori across LoRA weights and training steps.

LoRA weights	Training steps				
	5k	**10k**	**15k**	**20k**	**23k**
1	0.439	0.398	0.395	0.435	0.396
2	0.484	0.598	0.543	0.571	0.595
3	1.170	1.303	1.257	1.378	1.463

Table 2. FAD↓ scores for Korean pop across LoRA weights and training steps.

To quantitatively compare the model's adaptation to pansori and Korean pop, we computed the Fréchet Audio Distance (FAD) [12] scores for each genre under different LoRA weights(rows in Tables) and training steps(columns in Tables). The results are summarized in Table 1 and Table 2.

For pansori, Table 1 shows that the lowest FAD is 0.423 at 23k steps with LoRA 2. LoRA 3 outperforms LoRA 1 up to 20k. For Korean pop, Table 2 indicates that lower FAD values generally occur with LoRA 1 and early checkpoints, while LoRA 3 yields higher FAD across steps.

Overall, pansori benefits from sufficient LoRA weights and training steps, whereas Korean pop tends to perform best with minimal adaptation. While Korean pop outperformed pansori in FAD scores at the lowest LoRA weight, its performance degraded significantly when the LoRA strength was increased.

3.2 Expert Feedback

To assess the quality of pansori generation, we consulted Jueun Lee, the original performer of the dataset and a professional gugak musician. According to her evaluation, the model captured vocal timbre, expressive ornamentation, and contextually appropriate segmentation when symbolic markers like *jangdan* and *aniri* were provided.

Although jangdan may sound sparse compared to pop drum accompaniment, it follows a well-defined rhythmic framework that the model did not successfully learn. However, the model often failed to maintain *jangdan* cycles, resulting in unstable timing and unclear rhythmic phrasing.

The model also struggled with *aniri*—narrative segments—as it attempted to handle them within a singing-oriented generation framework. While the resulting *aniri* passages were distinct from sung segments, they still felt awkward and unnatural, failing to reflect their theatrical and speech-like nature. We provide audio examples on our supplementary webpage [1] .

4. CONCLUSION

FAD results suggest that the further a genre is from the model's pretraining distribution, the harder it is to steer generation through fine-tuning.

According to expert evaluation, the model captured short-term characteristics of pansori—such as vocal texture, ornamentation, and segment transitions—but failed to reproduce long-term characteristics of pansori-such as melodic progression and rhythmic cycles. *Aniri* lacked the natural theatrical quality expected of spoken narrative. These issues reflect the model's difficulty in handling the mixed characteristics of pansori, which includes both singing and speech-like narration. These findings imply that adaptation to traditional genres like pansori demands more than just parameter tuning. Rhythm-aware modeling, symbolic guidance, or architectural modifications may be necessary for more faithful generation.

5. ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-00560548)

6. REFERENCES

[1] "Suno," https://suno.com, accessed: 2025-08-29.

[2] "Udio," https://udio.com, accessed: 2025-08-29.

[3] S. Forsgren and H. Martiros, "Riffusion - Stable diffusion for real-time music generation," 2022. [Online]. Available: https://riffusion.com/about

[4] R. Yuan, H. Lin, S. Guo, G. Zhang, J. Pan, Y. Zang, H. Liu, Y. Liang, W. Ma, X. Du, X. Du, Z. Ye, T. Zheng, Y. Ma, M. Liu, Z. Tian, Z. Zhou, L. Xue,

[1] https://jarammm.github.io/pansorigen/

X. Qu, Y. Li, S. Wu, T. Shen, Z. Ma, J. Zhan, C. Wang, Y. Wang, X. Chi, X. Zhang, Z. Yang, X. Wang, S. Liu, L. Mei, P. Li, J. Wang, J. Yu, G. Pang, X. Li, Z. Wang, X. Zhou, L. Yu, E. Benetos, Y. Chen, C. Lin, X. Chen, G. Xia, Z. Zhang, C. Zhang, W. Chen, X. Zhou, X. Qiu, R. Dannenberg, J. Liu, J. Yang, W. Huang, W. Xue, X. Tan, and Y. Guo, "YuE: Scaling Open Foundation Models for Long-Form Music Generation," 2025. [Online]. Available: https://arxiv.org/abs/2503.08638

[5] Z. Liu, S. Ding, Z. Zhang, X. Dong, P. Zhang, Y. Zang, Y. Cao, D. Lin, and J. Wang, "SongGen: A Single Stage Auto-regressive Transformer for Text-to-Song Generation," 2025. [Online]. Available: https://arxiv.org/abs/2502.13128

[6] Z. Ning, H. Chen, Y. Jiang, C. Hao, G. Ma, S. Wang, J. Yao, and L. Xie, "DiffRhythm: Blazingly Fast and Embarrassingly Simple End-to-End Full-Length Song Generation with Latent Diffusion," 2025. [Online]. Available: https://arxiv.org/abs/2503.01183

[7] J. Gong, S. Zhao, S. Wang, S. Xu, and J. Guo, "ACE-Step: A Step Towards Music Generation Foundation Model," 2025. [Online]. Available: https://arxiv.org/abs/2506.00045

[8] E. Kanhov, A.-K. Kaila, and B. L. Sturm, "Innovation, data colonialism and ethics: critical reflections on the impacts of AI on Irish traditional music," *Journal of New Music Research*, vol. 53, no. 1-2, pp. 47–63, 2024.

[9] A. Mehta, S. Chauhan, A. Djanibekov, A. Kulkarni, G. Xia, and M. Choudhury, "Music for All: Representational Bias and Cross-Cultural Adaptability of Music Generation Models," *arXiv preprint arXiv:2502.07328*, 2025.

[10] Y.-S. Lee and C. ho Kim, Eds., *Korean Musicology Series, 2: Pansori*. Seoul: The National Center for Korean Traditional Performing Arts, 2008. [Online]. Available: https://www.gugak.go.kr/site/program/board/basicboard/view?currentpage=2&menuid=001003002005&searchselect=&searchword=&pagesize=10&boardtypeid=24&boardid=1409&lang=ko

[11] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," 2021. [Online]. Available: https://arxiv.org/abs/2106.09685

[12] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms," 2019. [Online]. Available: https://arxiv.org/abs/1812.08466