International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 683-686, May-2026

https://doi.org/10.47001/IRJIET/2026.105091

Timbre Transfer from Flute to Sarangi Using Latent Diffusion Bridge

[1]Aashish Shrestha, [2]Sanjivan Satyal

[1,2]Department of Electronics and Computer Engineering, Pulchowk Campus, Institute of Engineering, Lalitpur, Nepal

***Abstract* - Timbre transfer aims to modify the timbral characteristics of audio while preserving key elements like melody and rhythm. Advances in diffusion-based models have yielded promising results in image and audio synthesis. However, their application to ethnic Nepali instruments remains largely unexplored. We explore an unsupervised method for timbre transfer in Sarangi using latent diffusion bridges. In our experiment, the flute model maps the input audio into its corresponding Gaussian prior, and the Sarangi model reconstructs the target audio from the Gaussian prior. The trained Sarangi model can be used both as a source and a target model. Experimental results demonstrate that the model successfully keeps the melodic structure while altering timbral qualities.**

Keywords: Audio synthesis, Latent Diffusion, Sarangi, Timbre Transfer.

I. INTRODUCTION

Timbre is the quality of sound that lets us differentiate between sounds that have the same pitch and volume. Timbre of a sound source is affected by overtones, attack/delay, and material of construction. We can tell the difference between musical instruments or voices by looking at these traits. Timbre transfer in music is concerned with changing an audio signal so that one instrument sounds like another. While doing this, its semantic content and melodic structure are preserved. This means changing the recording of an instrument so that it sounds like it was played by a different instrument, without changing the main meaning of the performance [1].

The Sarangi is an ethnic Nepali musical instrument that has not been widely studied in audio research. It is a four-stringed bowed instrument traditionally played by the Gandharva community. It is known for its similarity to the human voice. Although Western instruments like the violin, guitar, and piano have been widely used for timbre transfer research, Nepali instruments such as the Sarangi remain largely unexplored. One of the main reasons is the lack of a high-quality, curated paired training dataset. In works like [2], they use synthetic data generated from audio synthesizers. This does not apply to Sarangi, which lacks such a synthesizer.

Like other audio generation tasks, timbre transfer methods work in one of the two domains: time or frequency. TimbreTron [3] used the Constant-Q Transform (CQT) along with CycleGAN [4] to perform unpaired translations between different musical instruments. These methods effectively capture the overall spectral structures; they often face issues with phase artifacts and lack fine temporal detail. Attention-based approaches like ATT [5] improved consistency by modeling long-range relationships in audio data.

Generative Adversarial Networks (GANs) struggled to generate locally coherent audio waveforms. GANSynth [6] demonstrated that GANs can be used to generate locally coherent, high-quality audio waveforms in the spectral domain. DDSP [7] combined classical signal processing with neural networks for audio synthesis. Their methods also allowed manipulation of each separate model component like pitch and loudness.

Diffusion-based generative models have become the leading approach in audio synthesis and transformation. DiffWave [8] and AudioLDM [9] have shown impressive results in high-quality audio synthesis through iterative noise reduction. DiffWave is a diffusion probabilistic model that converts white noise into a structured waveform using a Markov chain. It supports conditional as well as unconditional generation. AudioLDM is a text-to-audio system trained on CLAP [10] embeddings. The generation is conditioned on the text embeddings. Latent diffusion models (LDMs) work in a compressed representation space, significantly boosting efficiency. Probability Flow Ordinary Differential Equations (PF-ODE) [11] allow mappings between distributions without needing paired data. This method is especially suitable for instruments like the Sarangi, where well-organized paired datasets may be limited or unavailable.

II. METHODOLOGY

We extend the work of [2] in an unsupervised timbre transfer framework specifically for the Sarangi instrument. The main advantage of their method is that each model can be trained independently on its own instrument's audio data. By integrating the EnCodec neural audio compressor [12] with a latent diffusion bridge, we train an unsupervised timbre

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 10, Issue 5, pp 683-686, May-2026
https://doi.org/10.47001/IRJIET/2026.105091

transfer model that can be used both as a source and a target bridge for high-quality timbre transfer.

2.1 Data Preprocessing

The Sarangi audio is ensured to match a specified sample rate of 24,000 Hz. If not, the audio is resampled accordingly. Each audio sample is segmented into 10-second samples with a stride of 3 seconds. The overlapping audio is used to overcome the dataset size limitation. The audio is converted into monophonic by averaging along the channel dimension. The audio samples are organized into subsets of 80% for training and 20% for validation.

2.2 Components

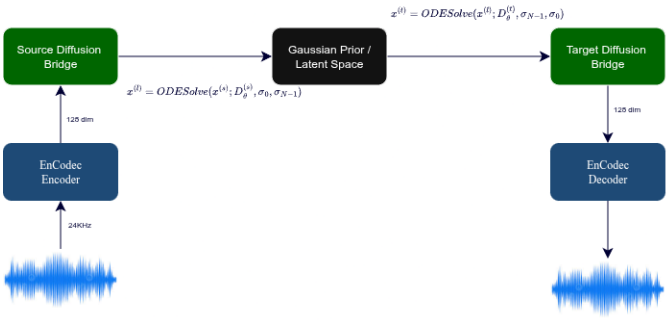


Figure 1: Components of end-to-end timbre transfer using Latent Diffusion Bridge. EnCodec converts the raw audio samples into a latent representation, and the source diffusion bridge adds noise; the target diffusion bridge removes noise and adds timbral characteristics of the target instrument. The latent is then converted back to an audio sample using the EnCodec decoder.

Two diffusion bridges, i.e., source and target, are required for end-to-end timbre transfer. The process begins by compressing the source audio into a latent representation via a pre-trained EnCodec encoder, which is then mapped to a shared Gaussian prior through a forward Probability Flow-Ordinary Differential Equations (PF-ODE) [11] using the source-specific diffusion model. This de-identification step effectively neutralizes the source timbre while preserving the underlying musical structure.

Subsequently, a target-specific diffusion model performs a reverse PF-ODE to map the noise from the prior into the target instrument's distribution. The EnCodec decoder reconstructs the transformed latents into the final target audio. The inference process can be summarized as follows:

$$x^{(l)} = ODESolve(x^{(s)}; D^{(s)}{}_{\theta}, \sigma_0, \sigma_{N-1}) \quad (1)$$

$$x^{(t)} = ODESolve(x^{(l)}; D^{(t)}{}_{\theta}, \sigma_{N-1}, \sigma_0) \quad (2)$$

where σ_{N-1} and σ_0 are the noise levels at step N-1 and 0, $D^{(s)}{}_{\theta}$ and $D^{(t)}{}_{\theta}$ are the source and target models. $x^{(s)} \epsilon R^d$ represent the data representation in the source domain. Similarly, $x^{(l)}$ and $x^{(t)}$ are the data representations in the latent and target domains. ODESolve is the Heun numerical solver as specified in [2]. It is theoretically shown that sequentially solving PF-ODE achieves cycle consistency, i.e., the process from source to latent and back to source recovers the original input [2].

2.3 Model Architecture

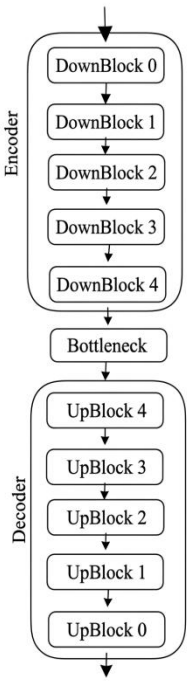


Figure 2: UNet architecture used as a denoiser. The input and output dimensions are shown alongside each block

The model architecture is based on a one-dimensional U-Net architecture that processes latent representations derived from EnCodec [12]. The encoder and decoder each consist of 5 blocks. The blocks of the encoder consist of one Conv1d layer and two ResNet blocks. This is followed by a transformer layer, except in the first block. This reduces the temporal dimension and expands the channels. In the decoder, there is one extra ResNet block at the attention level to accommodate the additional skip from the transformer output. Upsampling is performed after each ResNet block.

2.4 Training Objective

To train a model $D^{(s)}{}_{\theta}$, the L_2-norm of the original (x_0) and target embedding $(D^{(s)}{}_{\theta}(x_i; \sigma_i))$ is minimized. Target embedding is the denoised estimate of x_0.

$$E\left[\lambda(\sigma_i) \,||\, x_0 - D^{(s)}{}_{\theta}(x_i;\, \sigma_i) \,||_2{}^2\right] \quad (3)$$

where $\lambda(\sigma_i)$ is the weighting parameter depending on the noise level [13].

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 10, Issue 5, pp 683-686, May-2026

https://doi.org/10.47001/IRJIET/2026.105091

2.5 Evaluation Metrics

The performance of timbre transfer is evaluated on the following criteria: melody preservation and audio quality. For melody preservation, we use Jaccard Distance (JD) and Dynamic Pitch Distance (DPD). JD considers only the set of notes present in audio, while DPD also considers temporal note alignment [2]. For audio quality or perceptual quality, we use Fréchet Audio Distance (FAD). The Fréchet distance between two multivariate normal distributions having means μ_X, μ_Y and covariance matrices Σ_X, Σ_Y is given by:

$$FAD = || \mu_X - \mu_Y ||_2^2 + tr\,(\Sigma_X + \Sigma_Y - 2\sqrt{\Sigma_X \Sigma_Y}) \quad (4)$$

where *tr(.)* is the matrix trace [14].

2.6 Experimental Setup

The experiment aimed to perform timbre transfer between the Sarangi and the flute. The Flute model, trained by [2], is used to perform the timbre transfer with our Sarangi model. A one-dimensional U-Net model with 302M trainable parameters is used. The model is trained on a single NVIDIA RTX 4090 GPU with 24GB VRAM for 1000 epochs. AdamW optimizer is used with a learning rate set to $10e^{-4}$ and a batch size of 16. First and second momentum parameters were set as follows: β1 = 0.95, β2 = 0.999. The dataset consisted of 1,973 audio samples consisting of raw Sarangi audio samples played on melodies of different Nepali songs. The diffusion was performed on 128-dimensional latent audio embeddings produced using EnCodec [12].

2.6 Loss Curves

Training

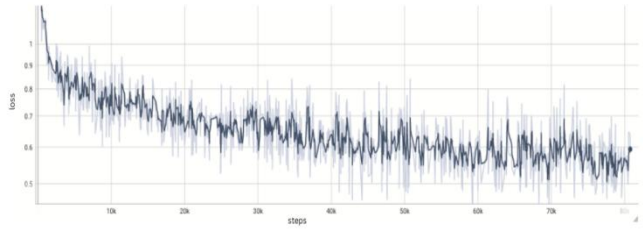


Figure 3: Loss over training steps. The final training loss is 0.58

Validation

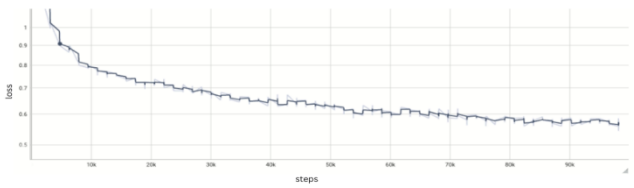


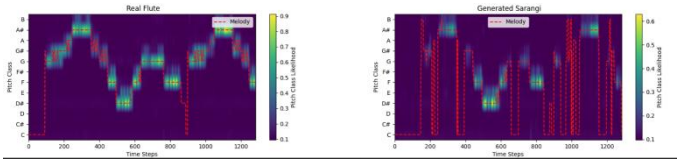
Figure 4: Loss over validation steps. The final validation loss is 0.57

III. RESULTS AND DISCUSSIONS

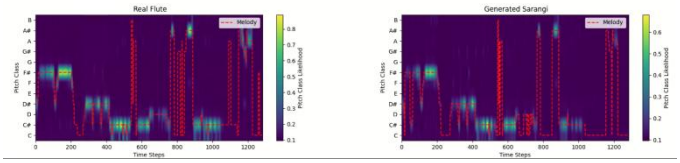
The performance of the timbre transfer is evaluated on melody preservation (DPD, JD) and audio/timbre quality (FAD). The model achieved an FAD score of 76.14, while DPD and JD values were 0.47 and 0.10, respectively. The qualitative analysis shows that there is clear evidence of melody preservation as well as timbre transfer.

Melodic Structure in Original vs. Generated Audio

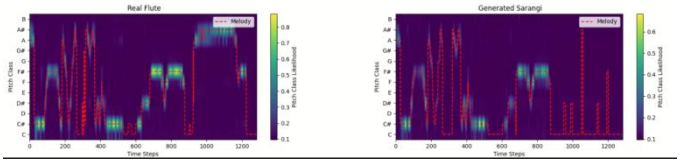
DPD: 0.76, JD: 0.1



DPD: 0.21, JD: 0.09



DPD: 0.61, JD: 0.18



IV. CONCLUSION

In this paper, we explore unsupervised timbre transfer in Sarangi using the latent diffusion bridge proposed by [2]. The research shows that the method is robust, and high-quality timbre transfer is achievable in Sarangi. However, our work is limited by the availability of Sarangi audio data. Future research can explore this method on a larger and more diverse training dataset, as well as extend it to support the generation of longer audio samples.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Department of Electronics & Computer Engineering, Pulchowk Campus, for providing a research opportunity and supporting us with computational resources. We would like to thank Kiran Nepali, the Founder of Project Sarangi, for generously providing raw audio samples for this project. We are indebted to all the Sarangi players, including Manice

ISSN (online): 2581-3048
Volume 10, Issue 5, pp 683-686, May-2026
https://doi.org/10.47001/IRJIET/2026.105091

Gandharva and Biraj Gandharva, whose performances were instrumental in training our models.

REFERENCES

[1] Bonnici, R. S., Benning, M., & Saitis, C. (2022, July). Timbre transfer with variational auto encoding and cycle-consistent adversarial networks. In *2022 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-8). IEEE.

[2] Mancusi, M., Halychanskyi, Y., Cheuk, K. W., Moliner, E., Lai, C. H., Uhlich, S., ... & Mitsufuji, Y. (2025, April). Latent diffusion bridges for unsupervised musical audio timbre transfer. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.

[3] Huang, S., Li, Q., Anil, C., Bao, X., Oore, S., & Grosse, R. B. (2018). Timbretron: A wavenet (cyclegan (cqt (audio))) pipeline for musical timbre transfer. *arXiv preprint arXiv:1811.09620*.

[4] Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232).

[5] Jain, D. K., Kumar, A., Cai, L., Singhal, S., & Kumar, V. (2020, July). ATT: Attention-based timbre transfer. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-6). IEEE.

[6] Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., & Roberts, A. (2019). Gansynth: Adversarial neural audio synthesis. *arXiv preprint arXiv:1902.08710*.

[7] Engel, J., Hantrakul, L., Gu, C., & Roberts, A. (2020). DDSP: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643*.

[8] Kong, Z., Ping, W., Huang, J., Zhao, K., & Catanzaro, B. (2020). Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*.

[9] Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., ... & Plumbley, M. D. (2023). Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*.

[10] Elizalde, B., Deshmukh, S., Al Ismail, M., & Wang, H. (2023, June). Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1-5). IEEE.

[11] Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

[12] Défossez, A., Copet, J., Synnaeve, G., & Adi, Y. (2022). High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

[13] Karras, T., Aittala, M., Aila, T., & Laine, S. (2022). Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, *35*, 26565-26577.

[14] Dowson, D. C., & Landau, B. (1982). The Fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, *12*(3), 450-455.

AUTHORS BIOGRAPHY

Aashish Shrestha,
Student, M.Sc. Computer Engineering, Specialization in Data Science & Analytics, Department of Electronics & Computer Engineering, IOE Pulchowk Campus, Lalitpur, Nepal.

Sanjivan Satyal,
Assistant Professor, Department of Electronics & Computer Engineering, IOE Pulchowk Campus, Lalitpur, Nepal.

Citation of this Article:

Aashish Shrestha, & Sanjivan Satyal. (2026). Timbre Transfer from Flute to Sarangi Using Latent Diffusion Bridge. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 10(5), 683-686. Article DOI https://doi.org/10.47001/IRJIET/2026.105091

****

