Sing Style Transfer: Navigating Cross-Genre Singing Style Transfer with VAE-GAN Architecture

Anonymous Author(s)
Affiliation
Address
email

Abstract

Voice style transfer has primarily focused on transforming speech from speaker to speaker, leaving singing style largely unexplored. To address this gap, we introduce SingStyleTransfer, a model utilizing a Variational Autoencoder-Generative Adversarial Network (VAE-GAN) architecture to perform style transfer across genres while maintaining the original vocal identity. Using the SingStyle111 dataset, the model has shown success in preserving audio semantic content and producing realistic genre-to-genre transformations. This project sets the stage for more advanced applications in music production, vocal coaching, and AI-driven performance synthesis, demonstrating its potential to personalize and adapt singing styles across various musical contexts.

1 Introduction

In recent years, research in Voice Style Transfer (VST) has gained significant attention, aiming at transforming one speaker's voice into another's. However, these advancements leave the realm of singing style relatively unexplored. In contrast to speech, singing involves melodic, rhythmic, and vocal nuances, and transferring these elements across genres presents a new and unique challenge. Despite advancements in voice conversion models like AutoVC [13] and HierVST [10], very few models aim at changing singing styles across genres while maintaining musical characteristics. To address this challenge, this paper introduces SingStyleTransfer: a model that maintains vocal characteristics while transferring genre-specific singing styles.

Singing style can be interpreted as the distinct characteristics associated with a song's genre that go beyond melodic and lyrical content. Variations in rhythm, phrasing, ornamentation, and emotion that are unique to specific genres such as pop, opera, or jazz, are independent of musical elements like melody and lyrics. For example, vocals in pop may be sung with a bright vocal tone, while the same music sung in an opera style may have a more resonant, full-bodied technique. To reflect this, we used the vocal dataset SingStyle111, which offers multilingual and multi-style monophonic voice recordings of professional singers. This dataset allows us to analyze distinct genre-specific styles in different performances of the same songs [3].

StyleSinger [17] is a key contribution to the field of singing voice synthesis, focusing on out-of-domain (OOD) style transfer [12] for generating high-quality singing voices with unseen styles. Although StyleSinger tackles challenges related to the expressive nuances of singing, such as timbre, emotion, and articulation, its dependence on reference samples for unknown styles limits its flexibility. Our approach directly transfers singing styles across genres by a hybrid Variational Autoencoder and Generative Adversarial Network (VAE-GAN) used previously for timbre transfer [1]. SingStyleTransfer allows for more generalized style transfer that preserves the original singer's melodic and lyrical content while adapting the stylistic nuances of the target genre. Moreover, due to the VAE-GAN's capability to preserve semantic content from style-specific modifications, we aim to

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

offer a more precise control over the performance characteristics transferred across genres to allow for seamless adaptation to new styles.

2 Data

The SingStyle111 dataset contains 111 songs sung by 8 professional singers totalling 12.8 hours of music [3]. In their dataset, around 80 songs cover at least two distinct singing styles performed by the same singer. Each song is distributed into 5-10 second samples and are studio recorded, monophonic, and created for the main purpose of style and voice conversions. In the multi-style recordings, singers were asked to exaggerate distinct performance styles to better highlight nuances between genres. For our purposes, we only extracted clips from the Pop, Opera, Rock, and Jazz genres as their songs were found to have the most overlap with each other.

Table 1: SingStyle111 Dataset statistics excluding data augmentations

Genre	Samples	Total Time	Average Sample Time
Pop	3605	6 hrs 40 min 46 s	6.67 seconds
Opera	1314	2 hrs 36 min 46 s	7.16 seconds
Rock	263	21 min 6 s	4.81 seconds
Jazz	321	35 min 45 s	6.68 seconds

To visualize the distribution and relationships between the different genres in our dataset, we created UMAP (Uniform Manifold Approximation and Projection) [11] plots of the extracted features from Pop, Opera, Rock, and Jazz samples split by male and female singers (Figure 2).

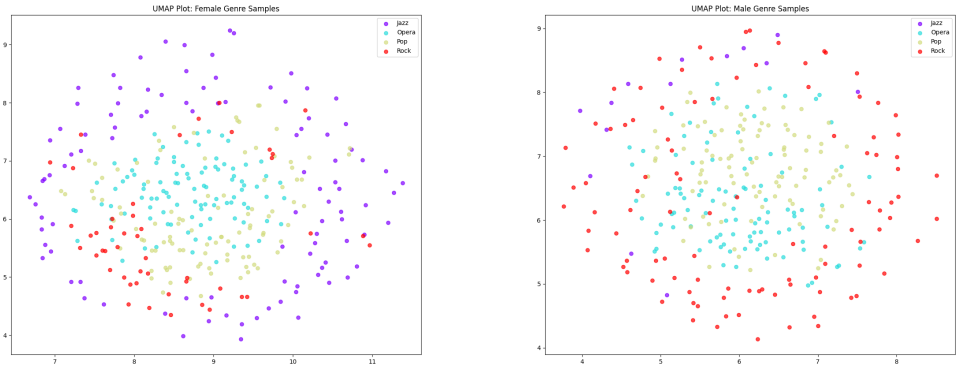

Figure 1: Female and male UMAP plots for Pop, Opera, Rock, and Jazz samples

In both plots, we observe partial separation between genres, demonstrating that the latent space representations capture genre-specific characteristics to some degree despite the samples differing in singer and song. The Rock and Jazz samples (red and purple) show wide dispersion and some overlap with other genres, suggesting a large spread of vocal styles possibly due to the improvisational nature and diverse influences in jazz vocals. Opera samples (light blue) form a more concentrated cluster, particularly in the female plot, which may reflect a more uniform singing technique in classical singing. Pop samples (light yellow) show moderate clustering with significant overlap with other genres, especially in the male plot, which could indicate the genre's tendency to incorporate elements from various musical styles.

3 Methodology

3.1 Data Preprocessing

Due to a lack of vocal variation in the SingStyle111 dataset (as it was limited to only 8 professional singers), we apply several data augmentation techniques, each with a 50% probability of being applied to any given sample. Augmentations include:

- Adding colored noise to simulate background noise variations and improve model robustness
- Multiplying the audio by a random amplitude factor, introducing variation in volume to make the model invariant to gain changes
- Pitch shifting the audio without affecting tempo to mimic different vocal ranges
- Flipping the audio samples' polarity, which can be beneficial when training models sensitive to phase differences
- Adding temporal variation without altering the content
- Reversing the audio along the time axis, creating variations in temporal structure similar to flipping an image

Since certain genres of the SingStyle111 dataset were deficient in samples compared to others, we duplicated tracks from combined genres, including Pop Jazz or Folk Rock, and assigned them to their corresponding individual genres (for example, Pop Jazz tracks were duplicated into both Pop and Jazz datasets). These augmentations were applied only to the training set to preserve the fidelity and authenticity of the validation and testing sets.

Similar to the paper, "Timbre Transfer with Variational Auto Encoding and Cycle-Consistent Adversarial Networks," the input audio is resampled to a sample rate of 16,000 Hz, normalized using root mean square normalization with a target amplitude of -30dB, and filtered to remove long silences [2]. Mel-spectrograms are then computed with 128 mel frequency bins and logarithmically scaled before being normalized with min-max scaling to accelerate training convergence, following a standard 80-10-10 split for training, validation, and testing [1].

3.2 Model Design

The design of our model utilizes the open-source implementations for the Variational Autoencoder-Generative Adversarial Network (VAE-GAN) and WaveNet Vocoder [2] used in a similar project on Timbre Transfer [1]. This VAE-GAN WaveNet vocoder structure allows the model to capture the essential features of the original singing voice while also generating realistic transformations into different styles. The VAE part of the model learns to compress the input voice into a smaller, manageable representation (latent space) and then reconstruct it back to its original form [9]. This process helps in preserving the core characteristics of the voice, such as melody and tone. The GAN consists of two components: a generator and a discriminator. The generator creates new audio samples based on the latent representation, aiming to mimic the style of the target genre while the discriminator evaluates these generated samples by processing the encoding through an Adam optimizer, ensuring they sound realistic and match the desired style [5]. This approach is particularly suitable for singing style transfer because it can adjust specific stylistic elements—like rhythm and ornamentation—without altering the fundamental vocal qualities.

After the GAN generates the transformed audio representation, a WaveNet vocoder is used to convert this representation back into a high-quality naturally sounding waveform. Vocoders function by synthesizing sound from abstracted, simplified audio features, and WaveNet is a particularly advanced vocoder known for producing exceptionally realistic and nuanced sound outputs [14]. WaveNet operates on a sample-by-sample basis, using autoregressive modeling to predict each subsequent audio sample based on the previous ones. This process allows the vocoder to capture intricate temporal dependencies, ensuring that the generated waveform maintains a high degree of naturalness and avoids common artifacts that can make synthesized voices sound robotic or artificial.

This structure is particularly suited for our singing style transfer purposes, as it maintains the musical integrity of the audio, transforming features in a way that only affects style, while also working well under generation and unsupervised training.

4 Results

We used a combination of two metrics, the Structural Similarity Index Measure (SSIM) [15] for reconstructions, and the Fréchet Audio Distance (FAD) [8] for translations. They were both evaluated on WaveNet Vocoded outputs.

Although SSIM is generally used for imaging, it allows for comparison of signal structures in spectrogram form. The SSIM metric is computed through gauges of luminance, contrast, and structural similarity, each comparing the signal's intensity, variation, and overall correlation, respectively. In the context of this project, SSIM evaluates how well the VAE-GAN model preserves the melodic and structural content of the original audio after passing through the reconstruction process. In comparison to other metrics like MSE (mean squared error) or PSNR (Peak Signal Noise Ratio), SSIM can better capture structural similarities in the spectrograms [16], which is essential for time-frequency representations. By applying SSIM to both one-pass and cyclic reconstructions found in Table 2, we assess how well the model maintains the integrity of the audio across these two stages, as maintaining the structural content of the melody and lyrics is crucial for successful style transfer.

FAD, on the other hand, evaluates the realism and perceptual quality of the final output audio. Similar to the Frechét Inception Distance (FID) [7] used in image generation tasks, the FAD adapts FID for audio generation through the use of embeddings generated from a VGGish model [6] that is pre-trained on a large-scale audio dataset called AudioSet [4]. In order to compare real audio samples to our model's generated ones, the FAD calculates the distance between the multivariate Gaussian distributions of VGGish embeddings between real and generated samples. This allows us to assess the converted transfer outputs without ground truth samples, making it ideal for evaluating the quality of the transferred styles across multiple genres.

Both the SSIM and FAD provide valuable insights to ensure that the core musical content (e.g., melody, lyrics) is preserved during the reconstruction process and gauge how well the model produces realistic, high-quality audio.

Genre Conversion	Average SSIM recon	Average SSIM cyclic	FAD
Pop to Opera	0.95	0.90	4.732753
Pop to Rock	0.95	0.89	6.983255
Pop to Jazz	0.95	0.90	2.247029
Opera to Pop	0.96	0.90	5.032680
Opera to Rock	0.96	0.89	13.010536
Opera to Jazz	0.96	0.90	4.212143
Rock to Pop	0.92	0.87	5.686240
Rock to Opera	0.92	0.87	10.295559
Rock to Jazz	0.92	0.87	5.410228
Jazz to Pop	0.94	0.89	2.639906
Jazz to Opera	0.94	0.89	4.363017
Jazz to Rock	0.94	0.88	7.049780

Table 2: Structural Similarity Index Measure (SSIM) and Fréchet Audio Distance (FAD) for each genre conversion

Figure 2 presents two mel spectrograms comparing a monophonic rock sample and our model's output of the same musical content transformed into an opera style. The operatic version shows more energy visible in the 75-100 mel range, potentially showing increased presence of higher harmonics prevalent in operatic voices. Furthermore, the converted spectrogram shows more pronounced horizontal undulations in sustained notes, particularly in the 25-75 mel range, indicating the model may have introduced or enhanced vibrato.

5 Discussion

Our experiments demonstrate that the VAE-GAN and WaveNet vocoder architecture successfully transfers stylistic elements between singing genres from the SingStyle111 without altering the core identity of the singer's voice.

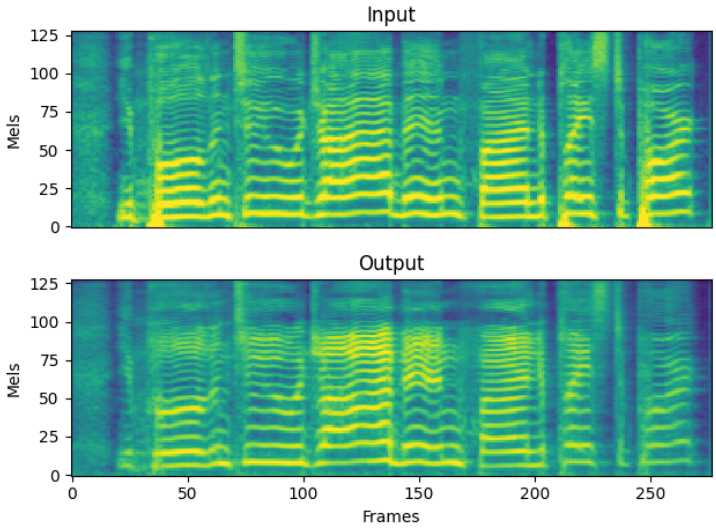

Figure 2: Mel spectrograms of a sample monophonic recording of the rock song "Your Mama Don't Dance" by Loggins & Messina as the input (above) and the output Griffin Lim reconstruction of its conversion to an Opera style (below).

The high SSIM scores across all genre conversion pairs (consistently above 0.87) suggest that the structural integrity of the original singing voice was preserved during the style transfer process. This is particularly evident in the genre pairs with high cyclic reconstruction SSIM scores, such as Pop to Opera and Opera to Jazz, where the original vocal characteristics remained intact after conversion and reconstruction to the original genre.

Lower FAD scores (between 1 and 5) indicated accurate representation with minimal degradation. Conversions from Pop to Jazz and Jazz to Pop yielded the lowest FAD scores, indicating high realism and minimal artifacts, while genre pairs like Opera to Rock resulted in higher FAD values. This difference in performances may be due to the more distinct characteristics between these genres, such as the more exaggerated dynamic range and timbral qualities of opera versus rock. Despite success in addressing data scarcity, the higher FAD values indicate that augmentations could not fully compensate for the lack of raw diversity and quantity in the training data.

6 Conclusion

In conclusion, the SingStyleTransfer model demonstrates the viability of using a VAE-GAN architecture for transferring singing styles across genres while preserving the original tonal qualities. By leveraging the SingStyle111 dataset, the model successfully transfers stylistic features in Pop, Opera, Rock, and Jazz, all while maintaining high fidelity to the original melodic and lyrical content. Structural Similarity Index Measure (SSIM) and Fréchet Audio Distance (FAD) metrics highlights the model's ability to balance content preservation and realistic style transformations.

While the current framework produces strong results, some conversions with higher FAD values indicate areas for improvement in generalization. Future improvements can focus on expanding the dataset to include more singers to increase the model's ability to generalize across different voices. Additionally, alternative vocoders or further tuning of augmentation techniques may improve the audio quality in more challenging genre pairs. Overall, SingStyleTransfer provides a flexible, genre-adaptive model that advances the field of singing style transfer and opens doors for applications in music production, vocal training, and entertainment.

References

[1] Russell Sammut Bonnici, Charalampos Saitis, and Martin Benning. Timbre transfer with variational auto encoding and cycle-consistent adversarial networks, 2021.

[2] Russell Sammut Bonnici, Charalampos Saitis, and Martin Benning. Timbre transfer with variational auto encoding and cycle-consistent adversarial networks, 2021.

[3] Shuqi Dai, Yuxuan Wu, Siqi Chen, Roy Huang, and Roger B. Dannenberg. Singstyle111: A multilingual singing dataset with style transfer. In *Proceedings of the 24th International Society for Music Information Retrieval Conference*, Milan, Italy, 2023.

[4] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017.

[5] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[6] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification, 2017.

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.

[8] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms, 2019.

[9] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[10] Sang-Hoon Lee, Ha-Yeong Choi, Hyung-Seok Oh, and Seong-Whan Lee. Hiervst: Hierarchical adaptive zero-shot voice style transfer, 2023.

[11] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[12] Toan Nguyen, Kien Do, Duc Thanh Nguyen, Bao Duong, and Thin Nguyen. Causal inference via style transfer for out-of-distribution generalisation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, volume 34 of *KDD '23*, page 1746–1757. ACM, August 2023.

[13] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, and Mark Hasegawa-Johnson. Autovc: Zero-shot voice style transfer with only autoencoder loss, 2019.

[14] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio, 2016.

[15] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[16] Zhou Wang and Alan C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009.

[17] Yu Zhang, Rongjie Huang, Ruiqi Li, JinZheng He, Yan Xia, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. Stylesinger: Style transfer for out-of-domain singing voice synthesis, 2024.