GENERATIVE SINGING STYLE TRANSFER ACROSS GENRES

Saanvi Bhargava[1,§] **Ethan Chu**[2,§] **Matthew Lee**[3,§] **Chuyang Chen**[4]
Kelvin Walls [4] **Bea Steers**[4] **Iran R. Roman**[5]

[1] The Harker School, San Jose, CA [2] Monta Vista High School, Cupertino, CA
[3] Riverdale Country School, Bronx, NY [4] New York University, New York, NY
[5] Queen Mary University of London, UK

§: equal contribution corresponding author: `i.roman@qmul.ac.uk`

Late-Breaking / Demo Session Extended Abstract, ISMIR 2024 Conference

ABSTRACT

Voice style transfer has focused on transforming speech between speakers, leaving singing style, independent of speaker, unexplored. We introduce SingStyleTransfer, a VAE-GAN to perform singing style transfer across genres. The model is evaluated on the SingStyle111 dataset for its ability to carry out genre-to-genre transformations.

1. INTRODUCTION

Research in Voice Style Transfer (VST) aims at transforming one speaker's voice into another's [1, 2]. However, these advancements leave the realm of singing style transfer relatively unexplored. In contrast to speech, singing style involves melodic, rhythmic, and vocal nuances unique to genres such as pop, opera, or jazz [3]. Transferring these elements across genres presents a complex challenge. Existing work includes StyleSinger [4], focusing on out-of-domain style transfer [5] for generating high-quality singing voices with unseen styles, and the work by Dai et al. (2018) [6], which aims to convert the singing style of a source singer into that of a target singer.

This late breaking demo introduces SingStyleTransfer. Using multilingual and multi-style recordings of professional singers, we trained a model to convert vocals from one genre to another. This means that, for example, the model converts a musical excerpt from a bright vocal pop style into a more resonant and full-bodied opera style. Our approach uses the hybrid Variational Autoencoder and Generative Adversarial Network (VAE-GAN) used previously for timbre transfer [7]. SingStyleTransfer allows for targeted style transfer, and leverages the VAE-GAN's capability to preserve semantic content from style-specific modifications. Our ultimate goal is to offer targeted control over the global characteristics transferred across genres.

2. DATASET

SingStyle111 contains 111 songs sung by 8 professional singers, totalling 12.8 hours [8]. In this dataset, around 80 songs cover at least two distinct singing styles performed by the same singer. Each song is cut into 5-10 second clips, and recorded for the purpose of style and voice conversions. Singers were asked to exaggerate distinct performance styles to better highlight nuances between genres.

3. METHODOLOGY

3.1 Data Pre-processing

Given its intended breadth and size, we seletively used some parts of the SingStyle111 dataset. First, we only use the tracks in the Pop, Opera, Rock, and Jazz genres. We split the audio tracks following a random 80-10-10 split for training, validation, and testing [7]. Data augmentation methods are applied to the training split, each with a 50% chance. Augmentations include: random scaling, polarity reversal, pitch shifting, time stretching, audio reversing, and adding colored background noise. Other data pre-processing remains the same as in Bonnici et al. [9].

3.2 Model Design

We use a hybrid VAE-GAN and WaveNet vocoder, based on the timbre transfer approach from [9]. We extended their model, which had two speaker categories (male and female), to four categories, corresponding to four musical genres: Pop, Opera, Rock, and Jazz. We train the model by encoding audio samples from the genres into the VAE, compressing them into a latent space [10]. The VAE decoder and GAN then generate spectrograms from these latent representations, with the discriminator refining their quality [11]. The WaveNet vocoder converts these spectrograms into natural-sounding waveforms by predicting each audio sample autoregressively [12]. 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 prioritizes musical integrity and realism under generation and unsupervised training.

Table 1. SSIM and FAD metrics on the test set by the VAE-GAN on SingStyle111 genre conversions, as well as the original voice conversion task (bottom two rows).

Conversion	SSIM Recon	SSIM Cyclic	FAD
Pop to Opera	0.95	0.90	4.41
Pop to Rock	0.95	0.89	6.82
Pop to Jazz	0.95	0.90	2.03
Opera to Pop	0.96	0.90	5.13
Opera to Rock	0.96	0.89	9.77
Opera to Jazz	0.96	0.90	3.91
Rock to Pop	0.92	0.87	5.68
Rock to Opera	0.92	0.87	10.30
Rock to Jazz	0.92	0.87	5.41
Jazz to Pop	0.94	0.89	2.67
Jazz to Opera	0.94	0.89	4.40
Jazz to Rock	0.94	0.88	7.04
Female to Male	0.89	0.80	1.65
Male to Female	0.87	0.73	2.96

4. RESULTS

We used a combination of two metrics: the Structural Similarity Index Measure (SSIM) [13] for reconstruction ("Recon" and "Cyclic" variants originally used by [7]), and the Fréchet Audio Distance (FAD) [14] for conversions. They were evaluated on Griffin-Lim reconstructions and WaveNet Vocoded outputs, respectively. SSIM and FAD each provide valuable insights to gauge how well the model produces realistic, high-quality audio, and to ensure that the core musical content (e.g. intonation, rhythm) is preserved during the reconstruction process.

4.1 Quantitative Results

The consistently high SSIM scores across conversions suggest that the structural integrity of the original singer's voice was preserved during the transfer. Moreover, lower FAD scores (between 1 and 5) indicated accurate representation with minimal degradation. Conversions between Pop, Opera, and Jazz performed well, with the lowest FAD scores. Conversions to and from Rock had higher FAD scores, possibly due to the genre's distinct vocal characteristics and data scarcity (only 4.8% of the dataset). Compared to the model's original performance in voice conversion [7], our higher SSIM scores suggest conversions that maintain the expected structure. However, higher FAD scores suggest a decrease in realism.

4.2 Qualitative Results

Fig. 1 compares melspectrograms of a Rock input song to the model, and the output in Opera style. The operatic version shows more energy visible in the 75-100 mel range, demonstrating increased presence of higher harmonics, which are expected and prevalent in the operatic style [15]. Fig. 2 visualizes the latent space of our VAE using UMAP dimensionality reduction. We provide separate plots for male and female singers in order to maximize clustering along genres and not the singer's gender. This visualization gives intuition about the model's relatively disentangled representation of the different genres.

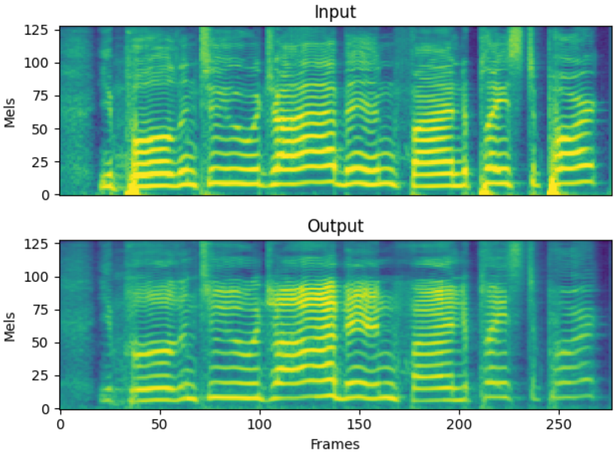

Figure 1. Mel spectrograms of "Your Mama Don't Dance" in Rock (input; top) converted to Opera (output; bottom).

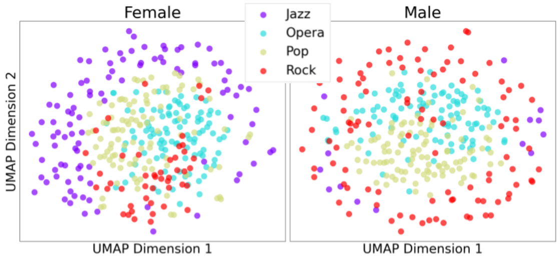

Figure 2. UMAP plot of the VAE's latent-space. Subplots separately show clips for female and male singers. Note the similar structure of the representational space in both singing genders (i.e. opera and pop towards the center versus jazz and rock forming an outer ring).

5. CONCLUSION

SingStyleTransfer demonstrates the viability of using a VAE-GAN to transfer singing styles across genres. Utilizing the SingStyle111 dataset, the model attempts to transfer features to and from Pop, Opera, and Jazz, all while aiming to maintain high fidelity to the original melodic and lyrical content. The SSIM and FAD metrics indicate the model's current ability to balance content preservation and realistic style transformations. While this work produces promising results, some conversions with higher FAD values indicate areas for improvement in generalization. Future improvements can focus on expanding the dataset in certain genres and 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 is a flexible, genre-adaptive model that explores the field of singing style transfer, with downstream applications in music production, vocal training, and entertainment.

6. REFERENCES

[1] K. Qian, Y. Zhang, S. Chang, X. Yang, and M. Hasegawa-Johnson, "Autovc: Zero-shot voice style transfer with only autoencoder loss," 2019. [Online]. Available: https://arxiv.org/abs/1905.05879

[2] S.-H. Lee, H.-Y. Choi, H.-S. Oh, and S.-W. Lee, "Hiervst: Hierarchical adaptive zero-shot voice style transfer," 2023. [Online]. Available: https://arxiv.org/abs/2307.16171

[3] E. Georgieva, P. Ripollés, and B. McFee, "The changing sound of music: An exploratory corpus study of vocal trends over time," in *Proceedings of the 25th International Society for Music Information Retrieval Conference*, San Francisco, California, USA, 2024.

[4] Y. Zhang, R. Huang, R. Li, J. He, Y. Xia, F. Chen, X. Duan, B. Huai, and Z. Zhao, "Stylesinger: Style transfer for out-of-domain singing voice synthesis," 2024. [Online]. Available: https://arxiv.org/abs/2312.10741

[5] T. Nguyen, K. Do, D. T. Nguyen, B. Duong, and T. 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*, ser. KDD '23, vol. 34. ACM, Aug. 2023, p. 1746–1757. [Online]. Available: http://dx.doi.org/10.1145/3580305.3599270

[6] C.-W. Wu, J.-Y. Liu, Y.-H. Yang, and J.-S. R. Jang, "Singing style transfer using cycle-consistent boundary equilibrium generative adversarial networks," 2018. [Online]. Available: https://arxiv.org/abs/1807.02254

[7] R. S. Bonnici, C. Saitis, and M. Benning, "Timbre transfer with variational auto encoding and cycle-consistent adversarial networks," 2021. [Online]. Available: https://arxiv.org/abs/2109.02096

[8] S. Dai, Y. Wu, S. Chen, R. Huang, and R. 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.

[9] R. S. Bonnici, C. Saitis, and M. Benning, "Timbre transfer with variational auto encoding and cycle-consistent adversarial networks," 2021.

[10] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2022. [Online]. Available: https://arxiv.org/abs/1312.6114

[11] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[12] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," 2016. [Online]. Available: https://arxiv.org/abs/1609.03499

[13] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[14] 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

[15] K. L. Reid, P. Davis, J. Oates, D. Cabrera, S. Ternström, M. Black, and J. Chapman, "The acoustic characteristics of professional opera singers performing in chorus versus solo mode," *Journal of Voice*, vol. 21, no. 1, pp. 35–45, 2007.