DMRN+16: Digital Music Research Network
One-day Workshop 2021

Queen Mary University of London
Tue 21 December 2021

Variational Auto Encoding and Cycle-Consistent Adversarial Networks for Timbre Transfer

Russell Sammut Bonnici*, Martin Benning, & Charalampos Saitis

Queen Mary University of London, United Kingdom, r.sammutbonnici@gmail.com

***Abstract*— The combination of Variational Autoencoders (VAE) with Generative Adversarial Networks (GAN) motivates meaningful representations of audio in the context of timbre transfer. This was applied to different datasets for transferring vocal timbre between speakers and musical timbre between instruments. Variations of the approach were trained and generalised performance was compared using the Structural Similarity Index and Frechét Audio Distance. Many-to-many style transfer was found to improve reconstructive performance over one-to-one style transfer.**

***Index Terms*—** Deep learning, Audio, Generative Adversarial Networks, Auto-encoders, Style Transfer, Timbre

I. Introduction

Timbre transfer is a task concerned with modifying audio signals such that their timbre is reformed while their semantic content is persisted. Through this, utterances of a speaker can be changed such that they sound like they were spoken by another speaker. Recordings of a source instrument can be manipulated in a similar way such that they sound like another target instrument played them. The challenge in making the modification take place first lies in how exactly timbral features can be captured.

II. Method

The approach adopted follows a UNIT inspired architecture that was initially proposed for voice conversion [1]. It uses a VAE for motivating content persistence that is embedded in a GAN for motivating timbre transfer. By applying this to the URMP dataset [2] for musical instruments, the generalisibility of the approach was challenged. An ablation study was also carried out on URMP and the Flickr 8k Audio dataset [3] for insight on what makes the architecture effective. Variations of the model included; a version with no Kullback–Leibler divergence (KLD) cyclic component for the VAE, a version where bottleneck residual blocks were used in place of basic residual blocks, and a version where the same model was trained for multiple style transfers at once (many-to-many) rather than one transfer (one-to-one).

III. Results

Table 1: Structural Similarity Index of Cyclic Reconstructions

Target	**Initial**	**No KLD Cyclic**	**Bottleneck Residual**	**Many to many**
Female 1	0.73	0.74	0.73	**0.77**
Male 1	0.80	0.78	0.68	**0.82**
Trumpet	0.83	0.83	0.78	**0.89**
Violin	0.81	0.81	0.78	**0.88**

Table 2: Frechét Audio Distance (General Vocoding)

Target	**Initial**	**No KLD Cyclic**	**Bottleneck Residual**	**Many to many**
Female 1	2.96	**2.77**	9.10	4.31
Male 1	1.65	2.48	6.97	**1.40**
Trumpet	**5.26**	5.52	6.06	5.85
Violin	**4.50**	5.52	12.68	4.99

The VAE-GAN approach was found general enough for applicability to instrument timbre transfer [4]. Basic residual blocks superseded bottleneck residual blocks around the latent space of the VAE for enriching content information. The presence of KLD for the cyclic loss component did not significantly impact performance. The many-to-many extension outperformed the initial one-to-one version in terms of reconstructive capabilities due to the increased variation of data passed through the universal encoder, yet improvements on the adversarial translation aspect were inconclusive. More clarity may be produced by training the utilised vocoder further.

IV. References

[1] E. A. AlBadawy and S. Lyu, “Voice Conversion Using Speech-to-Speech Neuro-Style Transfer,” in *Proc. Interspeech 2020*, 2020, pp. 4726–4730.

[2] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, “Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications,” *IEEE Transactions on Multimedia*, vol. 21, no. 2, pp. 522–535, 2019.

[3] D. Harwath and J. Glass, “Deep multimodal semantic embeddings for speech and images,” in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 237–244.

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

*Research supported by ENDEAVOUR Scholarships Scheme (Malta)