Queen Mary University of London

MSc.
Artificial Intelligence

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

Russell Sammut Bonnici

Co-Supervisor: Martin Benning
Supervisor: Charalampos Saitis

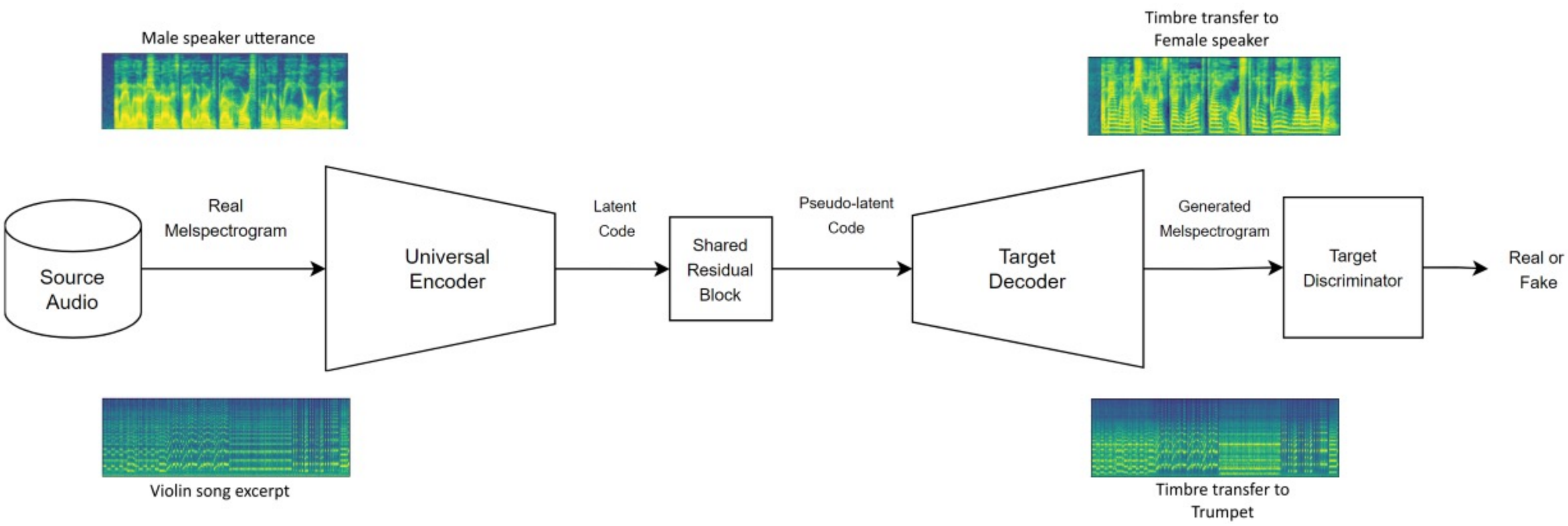
AIM

This research aims to investigate a generative deep learning approach for timbre transfer and prove its generalisability across timbral contexts of musical instruments as well as speakers. A quantitive evaluation across timbral contexts and model design scenarios is proceeded with for evidence on what contributes to a better timbre transfer.

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 to sound like another target instrument was played. The challenge in making the modification take place first lies in how exactly timbral features can be captured.

ARCHITECTURE



METHODOLOGY

The approach adopted follows a UNIT inspired architecture that was initially proposed for voice conversion. It uses a VAE for motivating content persistence that is embedded in a GAN for motivating style transfer. All VAE-GAN transfers were carried out in a mel-spectrogram format, and a WaveNet vocoder was used for converting results back into audio.
An ablation study was carried out on the URMP and the Flickr 8k Audio dataset for insight on what makes the architecture effective. Variations of the model included; a version with no Kullback–Leibler divergence (KLD) included in the cyclic reconstruction component, a version where bottleneck residual blocks were used in place of basic residual blocks, and one where the same model was trained for multiple style transfers at once (many-to-many) rather than one transfer (one-to-one).

CONCLUSIONS AND FUTURE WORK

In conclusion, the adopted VAE-GAN approach was able to transfer the timbre of musical instruments as well as speakers. It achieved an audible quality with a relatively inexpensive model working in the time-frequency domain. In the future, this model may also be applicable to the transfer of polyphonic timbre since it does not depend on a monophonic pitch transcriptions. The lack of a dependence on a monophonic pitch transcription likely hurts the quality for instrument timbre transfer, but at least this allows the approach to be general enough for application to more than just one type of audio style transfer problem.
A number of indications were found from the quantitative evaluation; basic residual blocks supersede bottleneck residual blocks around the bottleneck for enriching content information, that the presence of KL divergence for the cyclic loss component does not significantly impact performance, and finally, that the many-to-many extension outperforms the initial one-to-one version in terms of reconstructive capabilities due to the increased variation of data passed through the universal encoder. Though many-to-many improves the reconstructive aspect of the model, improvements on the adversarial translation aspect were not shown as per the Fréchet Audio Distance results. More clarity may be produced by training WaveNet with a more balanced dataset and with more data, or by adopting a different time-frequency vocoder with less sensitivity to data quantity.

RESULTS

TABLE III: SSIM of Cyclic Reconstructions

Dataset	Target	Model Experiments			
		Initial	No KLD Cyclic	Bottleneck Residual	Many to Many
Flickr	Female 1	0.73	0.74	0.73	**0.77**
	Male 1	0.80	0.78	0.68	**0.82**
	Female 2	0.76	0.76	0.66	**0.78**
	Male 2	0.70	0.70	0.57	**0.77**
URMP	Trumpet	0.83	0.83	0.78	**0.89**
	Violin	0.81	0.81	0.78	**0.88**
	Flute	0.64	0.65	0.62	**0.82**
	Cello	0.86	0.86	0.66	**0.91**

TABLE IV: Frechét Audio Distance (General Vocoding)

Dataset	Target	Model Experiments			
		Initial	No KLD Cyclic	Bottleneck Residual	Many to Many
Flickr	Female 1	2.96	**2.77**	9.10	4.31
	Male 1	1.65	2.48	6.97	**1.40**
	Female 2	**2.35**	**2.35**	8.04	2.64
	Male 2	**1.82**	1.95	7.03	2.90
URMP	Trumpet	**5.26**	5.52	6.06	5.85
	Violin	**4.50**	5.52	12.68	4.99
	Flute	**4.37**	4.40	6.39	5.64
	Cello	20.53	18.20	16.70	**16.21**

REFERENCES

1. M.-Y. Liu, T. Breuel, and J. Kautz, “Unsupervised image-to-image translation networks,” Advances in Neural Information Processing Systems, vol. 30, 2017.
2. E. A. AlBadawy and S. Lyu, “Voice Conversion Using Speech-to-Speech Neuro-Style Transfer,” in Proc. Interspeech 2020, 2020, pp. 4726–4730.
3. Sad 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.
4. 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.