Network Bending Neural Vocoders

Louis McCallum
Creative Computing Institute
University of the Arts, London
London, Peckham
l.mccallum@arts.ac.uk

Matthew Yee-King
Computing Dept.
Goldsmiths, University of London
London, New Cross
m.yee-king@gold.ac.uk

Abstract

Network bending [1] aims to elicit interesting creative output from generative neural networks by applying various transformations to the activations of groups of network nodes. This paper describes the investigation of how this emerging technique of 'network bending' can be used to provide novel creative control over sound synthesis networks based on the Magenta DDSP API [2] and how best to provide access to the resulting sound synthesis neural networks to creative practitioners.

1 Introduction

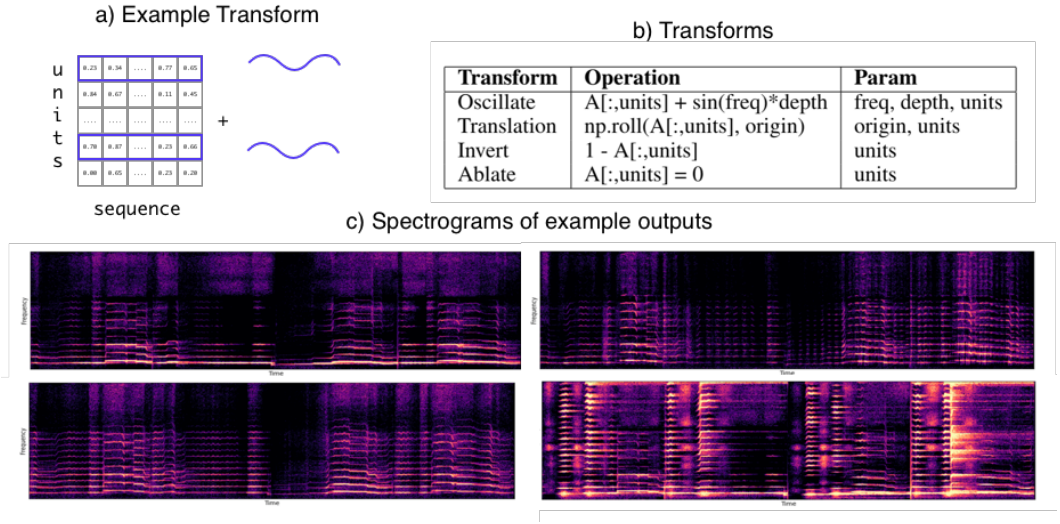
Google Magenta's DDSP API allows for rich audio synthesis to be controlled by deep learning models [2]. One of the primary use cases stated by its developers is timbre transfer, where an input audio signal is used as a control signal via audio analysis techniques. We are interested in how we might further control the output of the network beyond those parameters provided at the input, i.e. pitch and amplitude. Broad et al. [1] have taken an approach of eliciting interesting creative output from generative neural networks by applying various transformations to the activations of groups of network nodes. In this paper, we describe initial explorations with the use of this "network bending" in context of DDSP models, including which transforms to apply, where in the network to apply these tranformations and the parameterisation of tranformations over time.

2 Implementation

Compared to Broad et al.'s work [1], both the output domain (music, as opposed to visual art) and the structure of the models (Autoencoder with recurrent layers as opposed to GAN with convolutional layers) are different. This means it is unlikely that the techniques used by Broad et al. will be able to be transferred verbatim. Whilst the activations of a group of units in StyleGAN2 can be seen as representations of visual features in the training set, the output matrices from layers in the DDSP model represent the activations of units in one dimension, and a sequence of activations over time in another dimension. This means transformations such as rotations and reflections applied to the whole activation matrix (such as those used by Broad et. al) are in effect smearing time information across the 4 second block that will be crucial to reconstruction, especially with the recurrent layers. As such they are unlikely to result in any satisfactory or controllable sonic outcomes.

We instead look to apply transforms that preserve some information in the time domain whilst stimulating interesting timbral outcomes. We have developed a toolkit to apply any stacked combination of ablation, invertion, translation in the unit axis (wrap around) and oscillation layers after the first fully connected layer, after the GRU layer or after the second fully connected layer (see Fig. 1). The composer is able to pick a proportion of units to apply these to (selected randomly), with the translation and oscillation layers having parameters that can themselves be altered over time with either linear ramps or low frequency oscillators (LFOs).

4th Workshop on Machine Learning for Creativity and Design at NeurIPS 2020, Vancouver, Canada.



Transform	Operation	Param
Oscillate	A[:,units] + sin(freq)*depth	freq, depth, units
Translation	np.roll(A[:,units], origin)	origin, units
Invert	1 - A[:,units]	units
Ablate	A[:,units] = 0	units

Figure 1: a) How the transforms are applied b) Table of transforms and parameters c) Top left, clockwise (1) Clean (2) FC1 Oscillation (ramp on freq and depth, 100% units) (3) FC2 oscillation (LFO on freq and depth, 70% units) (4) FC1 Ablation (82% units)

3 Results

Fig. 1 demonstrates the spectral output (see here for audio examples [1]) of several uses of network bending using a model trained on approximately 12 minutes of Whitney Houston acapella tracks. A short excerpt from a different vocal piece was used as input. Our early explorations have shown that altering activations along the time axis, either through oscillating the activations themselves or altering parameters of the transformation layers over time, results in audible changes to the model output in a recognisable manner. We also see that adjusting the number of units the transforms are applied to alters the magnitude of the effects.

Primarily, we demonstrate how adding an oscillating signal with a linear ramp on the frequency to the activations of the first dense layer produces clear alterations to the audio output in the time and frequency domains in line with the increasing frequency of the control signal. This is also the case in the second dense layer, although we use a significantly lower depth of LFO (0.2 as opposed to 2.5) to avoid distorted, noisy whiteouts in the output audio. Finally, we show how ablating (zeroing) a proportion of units in the first dense layer creates a vibrato effect on the audio, with the influence of this effect proportional to the number of units transformed.

By exposing controllable parameters that are already in the electronic musician's toolkit, for example, LFO frequency and depth and parameter envelopes, we have been able to provide a workflow and interface that a computer musician will recognise from experimenting with and chaining together existing digital audio effect units. However, with potentially more exciting, non-standard timbral outcomes.

4 Future Work

Whilst at an early stage, our exploratory work opens up avenues in audio performance with DDSP models which may also be applicable to other models. This being said, even within the audio domain, different architechtures will generate different responses and it would be enlightening to combine our approaches with, for example, work by Collins et al. [3] that augments activations in spectral LSTM models. Moreover, we are currently using a fairly blunt, non-deterministic method of selecting sets of units and clearly, improved results will come from a more developed selection process. There are techniques for network interpretability that will prove useful in clustering units in this regard and work in the space of music generation has already been attempted by Brink [4].

[1] http://louismccallum.com/network-bending-audio-examples

5 Ethical Considerations

In our research, we take the position that AI and ML technology should be used to enhance and enrich human creativity and experiences. Our aim is not to replace the activity of human creatives with AI/ML systems. We would rather work towards AI tools that might be seen as supportive collaborators. Therefore, a lot of our work revolves around taking existing AI/ML systems and adapting them so that they can be accessed and manipulated by people who are not ML experts, but who do have creative domain expertise.

With specific regards to network bending, beyond displeasing machine learning devotees who may baulk at the subversion of a carefully optimised model, Collins et al. [3] evoke the spectre of brainwashing when rewiring the brains of AIs. They do however note this will not be a consideration until much more sophisticiated models exist, meaning our current interventions for creative goals are probably acceptable.

Acknowledgments and Disclosure of Funding

This work was supported by the Arts and Humanities Research Council funded project Musically Intelligent Machines Interacting Creatively (MIMIC) (mimicproject.com) and UKRI Human Data Interaction Network in collaboration with composer Max de Wardener.

References

[1] Broad, T., Leymarie, F.F. & Grierson, M. (2020) Network Bending: Manipulating The Inner Representations of Deep Generative Models, *arXiv:2005.12420 [cs.Cv].*

[2] Engel, J., Lamtharn, H., Gu, C. & Roberts, A. (2020) DDSP: Differentiable Digital Signal Processing, *in Proc. International Conference on Learning Representations 2020*

[3] Collins, N., Ruzicka, V., & Grierson, M. (2020) Remixing AIs: mind swaps, hybrainity, and splicing musical models, *in Proc. 2020 Joint Conference on AI Music Creativity*

[4] Brink, P. (2019) Dissection of a Generative Network for Music Composition, *Masters Thesis, School of Electrical Engineering and Computer Science*