University of Glasgow | School of Computing Science

A Deep Learning Approach to Musical Effects

Kieran McCool

School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
G12 8QQ

Level 4 Project — March 27, 2018

Contents

Chapter 1

Introduction

1.1 Musical Effects

Musical effects are transformations which can be applied to an audio track to change the sound in some way. These can take the form of simple tweaks to the frequency range of the track to applying pitch modulation and beyond.

Musical effects can completely change the way a track sounds and the art of mixing a track is - in many ways - more complicated than the composition and performance of the track. Sound engineers have a huge range of choices and responsibility when it comes to getting the soundscape correct for the final product.

The music industry is one of the few domains which is still to fully embrace digital technology. Many recording studios and musicians still make use of analogue equipment such as tape, vinyl, and vacuum tubes. As these technologies become more outdated and niche, the cost of maintaining and replacing equipment rises. As such, it is essential that software modelling catches up to the performance and quality of these older technologies.

1.2 Musical Effect Modelling

The main standard for producing digital effects for use with audio is through the Virtual Studio Technology (VST) protocol. This protocol defines a standardised definition for how a VST host would behave, allowing VST Plugins to be created which manipulate or use the audio track in some way.

The technology itself is fairly flexible and allows everything from audio visualisation tools to CPU intensive transformations to be applied. These effects can also be stacked and ordered easily and intuitively.

Most effects however, are limited to digital signal processing (DSP) based transformations, the likes of which are unable to replicate the non-linearity of many analogue methods. This is cause for concern in the music industry and is the reason many professionals and hobbyists claim that they are inadequate for anything more than practice and experimentation.

1.3 Machine Learning

In recent years, a huge amount of research has gone into machine learning. Its application frequently makes news for revolutionising the various industries it is applied to. As such, it would be interesting to see the effectiveness of machine learning in this field.

Of particular interest to us were convolutional neural networks and Long-Short Term Memory (LSTM) networks, which both seem to be at the forefront of deep learning research. However, the application of machine learning to audio data is in its infancy, resulting in quite a challenging implementation with many layers to it and several dependencies.

The idea of the project is to create two audio tracks, one which is a clean signal from an instrument and one which is the result of applying a VST effect to that clean track. We would then create a neural network which would take samples from the clean signal and be trained to produce the value of the corresponding sample in the processed track.

Chapter 2

Background

2.1 Nebula VST

Nebula yields the best results current modelling technologies can achieve. It achieves this using Volterra Kernel Modelling to better represent non-linear behaviours and exhibiting some level of memory capacity. This allows it to model time-based effects and non-linearity more effectively than primitive DSP approaches.

In doing this, it has become one of the most praised pieces of modelling software, criticised only for its system requirements and its steep learning curve. Requiring a minimum of 8GB of RAM with a recommended amount of 16-128GB, it is a rather demanding piece of software.

Given that Nebula is the best plugin available at this time, for this project to be considered a success, it should be as good as or better than Nebula.

2.2 Project Magenta

Project Magenta is a venture by the Google Brain Team, their goal is to explore the use of machine learning in creating art, with a particular interest in music. Most of their research is in audio synthesis with some ventures into genre classification and similar problems.

However, this project does raise some interesting questions about the ability of deep learning to characterise the complexities of music.

Most of their work however, was limited to interacting with MIDI data rather than with raw audio samples, perhaps limiting the complexity of the problem, making it an unfair comparison in terms of complexity to effect modelling.

2.3 WaveNet

Another use of deep learning for audio synthesis is in the field of generating realistic text-to-speech. DeepMind's WaveNet is a generative network, trained using raw audio samples to predict the value of the next sample in the series.

The project found that while generally an LSTM is better suited to this kind of time-series problem, they could achieve similar results using stacked convolutional layers applying dilated casual convolutions. At each layer they would double the dilation amount up to a maximum before starting again at 1. This a increases the receptive field of the network in a similar way to how an LSTM selectively remembers traits while maintaining the ease of training of a convolutional network.

Another way in which the WaveNet implementation differs from a standard convolutional network is in its output layer. Instead of predicting a sample as a floating point value representing the amplitude, the output is a one-hot encoded vector where the largest index corresponds with the μ–Law encoded integer of the sample.

$$F(x) = sgn(x)\frac{ln(1+\mu|x|)}{ln(1+\mu)} - 1 <= x <= 1$$

μ Law as expressed in SOURCE THIS

This network has proven itself to be capable of generating extremely effective, almost human sounding text-to-speech and is far easier to train than a standard LSTM-based approach.

Chapter 3

Aims

3.1 Effect Choices

The aim of this project is to explore the viability and success of deep learning in effect modelling at a fairly general level. As such, a multitude of different effects should be tested.

While there is a huge range of commonly used effects, and an even larger range of different variations of these same effects with subtly different responses, most of them fit in to one of three groups.

Amplitude Based	Frequency Based	Time Based
Distortion	High/Low Pass Filter	Chorus
Fuzz	Pitch Shift/Octave	Delay
Compression		Reverb
Amplifier/Cab Simulation		

3.2 Deep Learning

- Growing use of machine learning across all manner of disciplines
  - Image recognition, speech synthesis, text-to-speech.
  - Why not audio signals?
- Project aims to investigate the viability and explore the limitations of machine learning as a method for modelling effects.
  - Success would involve being able to convincingly reproduce the characteristics of a given effect after being trained on a clean input and the processed signal.
  - If the network can run over unseen data and produce the effect reliably.
  - Different measures of effectiveness, but more on this in evaluation. Perhaps introduce here though.

Sections below are unfinished

Chapter 4

Methods

- Network Architecture
- Evaluating outputs, Mel Spectogram etc

Chapter 5

Implementation

- PyTorch
- SciPy.signal
- LibRosa
- Reaper and the VSTs used
- How the pieces fit together

Chapter 6

Results

- Split into quantitative and qualitative
- Signal Analysis (Quantitative)
- Spectogram GIFs (Quantitative)
- Loss over time / Standard Deviation of loss (??? - Loss hasn't really proven to be particularly meaningful)
- ABX tests with classmates/friends (Qualitative)

Chapter 7

Conclusion

Chapter 8

References