Deep Audio Style Transfer:
Dynamics and Humanization In Sound Synthesis

Jacob Potter

Introduction

The goal of this project is to apply realistic dynamics and imperfections to digitally synthesized sound sources. Input sounds (synthesized) will be automatically identified as belonging to some combination of staccato, legato, pizzicato, tremolo, or chord playing styles. The program will then generate a new waveform which has a similar spectral signature and timbre, but with realistic dynamics and ‘human’ imperfections which match the respective style. Note that this does not aim to recreate real instruments, but to bring 'realism' to any arbitrary digitally synthesized sound.

Previous Solutions

Engineers have been trying to humanize synthesizers for decades. This is usually accomplished by applying randomized variations to each note such as modified volume, timbre, noise, and tempo. One issue with this is that these parameters do not actually behave randomly, but are governed by complex variables. Analog synthesizers are sometimes viewed as superior because voltage variations provide intrinsic imperfection and ‘realness’ which contributes warmth, depth, and grain. Many professionals agree that modern digital synth has closed the gap and can simulate any of the previously held advantages of analog synth, while others believe it will never be possible to completely remove the disparity. By applying a CNN to this task we can take the guesswork out and achieve a new level of realism matched only by instruments themselves.

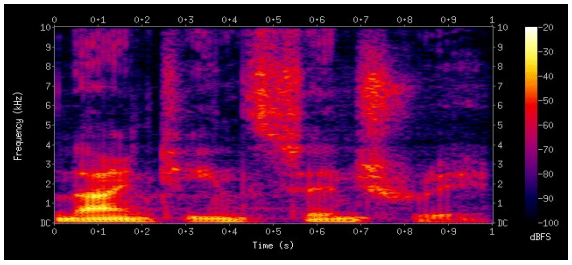
Training

Training data will consist of real recordings of instruments labeled with their respective playing style. Sound sources will be converted from waveform to frequency domain using the FFT algorithm. The CNN may benefit from normalizing the fundamental pitch of all samples to an arbitrary constant, such as 100 hrtz. Placing the fundamental in the same place for every sample could simplify things and allow for better perception of the dynamics. However, playing higher notes on some instruments can affect the timbre, which may be a desirable characteristic to transfer. It may be worthwhile to train on both sets with normalized and non-normalized pitch separately and see which has a better outcome. The data will then be written to spectrogram images which will be directly used for training.

Methods

The program will not only have to learn to identify playing styles, but also learn the musical inflections and dynamics associated with each style, and be able to apply those inflections to any arbitrary signal. Identifying the style of a signal should be trivial. It is only a matter of the envelope and the pitch of the fundamental frequencies, both of which are easy to see in a spectrogram. Applying the dynamics can be accomplished by reconstructing the spectrogram so that the frequencies at a given time are similar but have subtle modifications. The modifications might include time stretching, noise, pitch variations, distortion, frequency boosting, and filtering.

Example of an Audio Signal Spectrogram



List of characteristics to replicate

Acoustics and Dynamics

- Forte vs Piano (loudness) and their effects on Timbre
- Ambience and Acoustics
- Reverb and Reflections
- Signal Distortion
- Phasing and Wave Interactions

Imperfections and Realism

- Imperfect Velocity
- Imperfect Timing
- String Noise
- Accidental Harmonics

Style:

- Damping/Applied Pressure
- String Bending and Sliding
- Harmonics
- Staccato, Pizzicato, Tremolo
- Swing/Time bending

Related Materials:

Audio style transfer using CNN's
Audio Style Transfer
Neural Style Transfer on Audio Signals | Intel® Software
[1710.11385] Audio style transfer

Audio style transfer using GAN's
https://towardsdatascience.com/voice-translation-and-audio-style-transfer-with-gans-b63d58f61854

Effects of loudness on dynamics
https://asa.scitation.org/doi/full/10.1121/1.3633687

"TimbreTron"
https://www.cs.toronto.edu/~huang/TimbreTron/index.html

Issues with using CNN's on spectrograms
https://towardsdatascience.com/whats-wrong-with-spectrograms-and-cnns-for-audio-processing-311377d7ccd

Instrument Classification
https://www.ijitee.org/wp-content/uploads/papers/v8i9/I7728078919.pdf

Datasets

OpenMIC Dataset
https://zenodo.org/record/1432913#.Xi4hTDJKjcu

More Data sets
https://www.audiocontentanalysis.org/data-sets/

Spectrogram Generator
http://spek.cc/

Terms

Staccato: Notes sharply separated, are followed by silence.
Legato: Smooth and liquid, no silence between notes.
Pizzicato: Plucking.
Tremolo: Rapid reiteration of a note.
Chord: Three or more notes played simultaneously.
Forte: Loud, hard.
Piano: Quiet, soft.
Analog Synth: Uses analog circuits and analog signals to generate sound electronically.
Spectrogram: Visual representation of sound where the X axis denotes time, the Y axis denotes frequency, and the color denotes the intensity of the frequency.
Timbre: The quality of a sound making it distinct from other sounds of the same pitch.
Velocity: Volume of a note.
Dynamics: Variations in velocity and timbre in each note.
Humanization: Imperfections and noise caused by human error.
Harmonics: The phasing and variance of overtones in a note.
Envelope: How a note changes with time, usually denoted by attack, decay, sustain and release.
Fundamental: The lowest frequency in a note, as distinct from the overtones, which determines its pitch.