MODELING OF THE ACOUSTIC SIGNAL OF AN ELECTRIC GUITAR AMPLIFIER USING RECURRENT NEURAL NETWORKS

A Thesis presented to the Faculty of the Graduate School

University of Missouri-Columbia

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science

---

by

JOHN C. H. BENJAMIN

Dr. Yunxin Zhao, Thesis Advisor

DECEMBER 2021

The undersigned, appointed by the Dean of the Graduate School, have examined the thesis entitled:

MODELING OF THE ACOUSTIC SIGNAL OF AN ELECTRIC GUITAR AMPLIFIER USING RECURRENT NEURAL NETWORKS

Presented by John C. H. Benjamin

A candidate for the degree of Master of Science

And hereby certify that in their opinion it is worthy of acceptance.

Prof. Yunxin Zhao, Chair

Prof. Guilherme DeSouza

Prof. Carolina Heredia

Prof. Grant Scott

ACKNOWLEDGEMENTS

My time as a graduate student at The University of Missouri-Columbia is a cherished memory. It was a foundational experience of my life and one I hope others may enjoy in the future.

I want to thank my best friends and my mother. My mother helped me the whole way with motivation and support. All of my friends in the Computer Science, Electrical Engineering, and Math departments who I learned and shared career interests at Mizzou. We are all over the world working for the most influential technology companies and business sectors. It is an honor to have learned with everyone.

I also want to thank all of the amazing faculty at the University of Missouri for helping me along the way: Dr. Matthew Klaric and Dr. Grant Scott at the Center for Geospatial Intelligence for practical work experience and a laboratory home away from home; Dr. Yi Shang for advising; Dr. William Harrison for research mentorship; Dr. DeSouza and Dr. Zhao for training on machine learning, the topic of this thesis. Dr. Carolina Heredia in the Department of Music for facilitating testing and presentation to students of music. And a special thanks again to Dr. Zhao as my advisor for helping me on a thesis topic of great personal interest.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DIGITAL MODELING OF THE ACOUSTIC SIGNAL TRANSFORMATION OF AN ELECTRIC GUITAR AMPLIFIER USING RECURRENT NEURAL NETWORKS

John C. H. Benjamin

Dr. Yunxin Zhao, Thesis Supervisor

ABSTRACT

Neural networks have topped performance measures across a wide variety of computational tasks. These performances are prevalent within the domain of human perception type tasks such as classification or generation of images, audio, or text. Recurrent neural networks are neural network architecture of choice for time-domain data such as spoken or written human language. Recurrent neural networks have also shown promise for tasks in the specialized signal domain of music. This thesis explores using recurrent neural networks to model the particular musical qualities generated by analog electronic musical equipment. The electric guitar amplifier is used by electric guitar players to shape the timbre of their musical instrument as they play. Most professionals consider analog amplifiers designed to provide acoustic distortion with vacuum tubes as having the best sound and feel for musicians. We attempted to model the sound transformation of a vacuum tube based electric guitar amplifier using a convolutional recurrent neural network architecture. For our experiment, we trained recurrent neural networks of various architectures using inputs of electric guitar signals and the subsequent signal processed through a typical vacuum-tube based amplifier and audio recording equipment. Training data was collected according to the recommended specifications of previous experiments. The amp simulation models were compared against the original amplifier with signal analysis and subjective listening tests. Sound recording techniques for capturing the best

input and output data for the current state-of-the-art were analyzed. Analysis of various model input configurations and hyperparameter settings also showed several major limitations in the ability of the model to accurately reproduce the acoustic properties of a signal chain with more complex distortion characteristics than those previously tested. The limitations were analyzed and features of a new model were proposed. Subjective listening tests with three expert musical listeners showed that even though listeners could identify the quality degradation of the model, some listeners found the signals provided by specific models to be of more interest to their musical tastes. These models tended to model an amplifier set to a smooth gentle tone, which the model made less harmonically rich. However, models trained on amplifier tones with dense harmonic distortion characteristics were uniformly judged as very poor quality. These models could not accurately reproduce the intense compression and non-linear distortion qualities and ended up sounding abrupt with a hissing buzz.

CHAPTER 1: INTRODUCTION

1.1 Using Recurrent Neural Networks to model musical harmonics and dynamics

The purpose of this thesis project is to explore the possibility of getting the best possible sound out of digital musical software by modelling analog audio equipment with artificial intelligence techniques. The specific context is about creating a component of an electrical musical instrument that is designed to shape the tone or timbre of that instrument as it is played. Our goal is to model this change in timbre using a machine learning framework for signals called recurrent neural networks. The task of the neural network is to accurately model the waveform for both its dynamics and harmonics across a range of possible inputs from an electric guitar.

1.2 Acoustic Signals and Digital Audio Processing

Most modern consumers have heard an electric guitar played on a popular music recording. The electric guitar is ubiquitous in the musical culture of mass media with a rich heritage of artistic and technological innovation. What you hear online or on the radio is not just the performance of the musician, but a chain of electronic and digital signal modification choices created by the production team. With today's technology, an individual alone on a personal computer has access to the tools necessary to create a professional sounding audio production. A musician may plug an electric guitar player directly into a peripheral attached to his home computer to record a performance. The musician may take the initial signal and modify or mix the performance using different amplifier, microphone, or other settings tailored to the music. Despite the convenience of digital modelling technologies, most audio engineers and musicians prefer the musical sound qualities of analog electronic components recorded in a professional manner. For the case of electric guitar, there are a selection of affordable digital modelling software plugins. For the serious electric guitar player you will more often find an expensive and

fragile analog amplifier and speaker cabinet setup being used. Modern machine learning techniques offer a promising new approach to capturing the sonic qualities of analog electronic equipment in a digital format. This project aims to build and test a digital implementation of a high-end vacuum tube based electric guitar amplifier using recurrent neural networks in order to see if they can come closer to the tonal quality of analog equipment within a digital recording environment.

1.3 Deep Learning Research for Various Multimedia

Advances in speed and availability of computational power have also enhanced research in the artificial intelligence technique of multilayer neural networks for a variety of media applications. Deep learning has proven effective in a variety of generation or classification tasks for images, video, text, audio, and other types of data. In the domain of music, neural networks have proven effective in classifying genres of music, generating musical compositions similar to composers or genres, and synthesizing new sounds from existing timbres.

Deep Learning is of particular interest to musical audio because of the inherent difficulty in programming musical features of audio signals by hand. Machine Learning has grown as a discipline in Artificial Intelligence because of researcher's realizing that allowing machines to learn complex patterns in data is usually a better approach than trying to program them by hand. This is because many human perception and generation tasks inherent in our biology turn out to be too complicated or not well understood enough to program by hand. This realization applies as well to music. Like other art media, what constitutes 'good' musical qualities in a signal transformation or composition is not easily ascertainable or programmable.

1.4 Accurate acoustic reproduction in digital music arrangement

Most of the research in deep learning related specifically to music has to deal with general issues of musicology and arrangement such as generating rhythms, melodies, or harmonies or classifying works based on similarity. Another interesting approach for machine learning among artists and composers is to speed up or streamline the production component of a recorded piece of music such as the mixing or mastering of a recording session. The approach of this thesis is to replace components of audio signal chains with machine learning algorithms in order to expedite and democratize the production possibilities for musical performance.

1.5 Project Motivation

This project aims to expand upon existing experiments testing recurrent neural networks in modeling electric guitar amplifiers. Unlike other experiments this project attempted to build a complete signal chain for working use with studio musicians and audio engineers to 'reamplify' a guitar track. The experiment tested a complete signal chain rather than isolated components of an electric guitar studio recording setup and evaluated for use with musicians.

An amplifier is an important component of the audio signal chain for a popular electrical instrument. The electric guitar is a ubiquitous and popular instrument. The amplifier is a necessary component to an electric guitar. Electric guitar amplifiers traditionally come in two varieties: solid state and tube. The essential difference is whether the gain stages of the amplifier use transistors or vacuum tubes to amplify an incoming audio signal. Vacuum tubes are widely used in high-end electric guitar amplifiers because of particular musical qualities associated with their response to a user's manipulation of an electric guitar. Vacuum tubes are expensive and fragile. They are rarely used outside of professional audio or musical applications. Both solid state and vacuum tube based amplifiers ('Tube Amps') have been modeled using digital signal algorithms. These so-called modelling amplifiers may come in hardware components based on

digital signal processing chips or simply be pieces of software. This thesis explores whether deep learning recurrent neural network models are suitable for modelling amplifiers or complete signal chains for home studio use.

1.6 Thesis Organization

This thesis is organized by discussing the background, the purpose, the hypothesis, and the testing of an implementation of a machine learning model. For the background, we cover topics of acoustic signals as well as what makes a signal musical. We then cover the conversion of acoustic signals to analog electrical and digital signals. Next, the specific case of an electric guitar and its amplification are discussed, including the musical properties of vacuum-tube based amplifiers.

After an initial background we discuss the technology of recurrent neural networks. First we talk about machine learning in general. Next we cover the background of multilayer neural networks. We then go into the design of recurrent neural networks for sequential data. Finally we discuss the current literature in applying recurrent neural networks to the specific case of music for creating timbre.

After a brief on the background of the problem and the solution technique we discuss the limitations of the model found through experimental use, as well as discussions of the possible usefulness of the current models for musical composition based on listener tests.

CHAPTER 2: DIGITAL AND ANALOG AUDIO SIGNALS

2.1 Sound and music concepts

Sound is defined as an acoustic vibration travelling through a fluid medium. Our ears pick up sound vibrations and make sense of the sound through our learned understanding of human meaning. A normal human ear is capable of hearing sound vibrations around the frequencies of 20 to 20,000 Hertz. Humans have developed the capacity for complex artistic expression through the medium of sound as music. [2]

Music is an art form organizing sounds as melodies, harmonies, and rhythms over time. Melodies and harmonies are made of tones that have a specific fundamental frequency. According to the Western musical tradition there are 12 tones within an octave. A tone with the same label within a different octave is related by a whole fraction. A tone of the same letter and incidental symbol one octave higher is twice the frequency of the scale lower. This implies that the tone frequencies of this system are organized according to a logarithmic scale. Melodic musical instruments, such as instruments which make a sound from a plucked string, are designed to allow a performer to create a sequence of sounds according to these tones.

2.2 Harmonics

A musical note is defined in pitch according to its fundamental frequency. This frequency as generated by a musical instrument comprises the lowest, most powerful frequency. A natural sound, however, is a mixture of many vibrations at various frequencies mixed together with the fundamental frequency. Musical instruments are designed to supplement the fundamental frequency of a played note with many other supplemental frequencies that are pleasing to the human ear. Each musical instrument has its own signature of different frequencies created

simultaneously with the fundamental frequency of a particular note. This signature of frequencies creates what is called the timbre of an instrument. [2]

The content of this musical timbre may be analytically described using harmonics. Harmonics represents the frequencies available in a sound wave in reference to the fundamental frequency, as well as the amplitude of that frequency in the waveform. For example, the second octave harmonic is half the wavelength (one octave up) from the fundamental frequency. Any type of waveform that is not a pure sine wave introduces harmonics to the sound spectrum. The human ear is particularly attuned to odd and even harmonics in music. Even harmonics create a pleasant or happy sound, while odd harmonics sound uneasy and ominous.

2.3 Audio Signals

In order to record a musical performance or broadcast, the pressure waveforms in the air must be transferred into an electrical signal in a circuit. The relative frequency and strength of the sound corresponds to the frequency and amplitude of the electrical signal. An electrical circuit has a fixed amount of current that can be carried at one time, so some form of amplification must be performed in order to translate back from electrical signal to sound. The translation of electrical signal to sound is usually done with a loudspeaker, a device that uses electromagnetic force to drive a cone shaped speaker to vibrate according to the signal. [4]

In the past, all recordings were done using acoustic instruments, human voice, or other authentic sound sources. These acoustic sources were translated to electrical signals by microphones. Microphones are specially designed to convert acoustic waveforms travelling across a single point into an electrical current. A quality microphone will accurately capture the relative dynamics and harmonics of a sound. [4]

Modern music recording utilizes other sound generation techniques within a recording. A musical sound may be created native to an analog or digital signal. Electric guitars, for example, generate an electric current by exciting a metal wire across a coil. Another possibility is analog synthesizers, which are devices used to generate analog or digital signals electronically. There are also many forms of digital sound synthesizers hosted by digital audio environments.

2.4 Digital Audio and Analog Signal Conversion

Digitally encoded signals are necessary for computers to make sense of audio. However digital information is discrete whereas sound and an electronic signal is continuous. Representation of an audio signal in a digital domain can be done in a variety of ways. Linear Pulse Code Modulation (PCM) is most commonly used. In a PCM encoding, the signal waveform is sampled at regular intervals and represented as a value, typically between 1.0 and -1.0, amplitude of the waveform. A PCM is linear when each digital value represents the same interval of time as all other values. The time resolution of a PCM can be measured in bits per second for the number of samples in a second. The bitwise resolution depends on what size of floating point representation is used for each time sample. Analog signals are converted to PCM using specialized converter chips. A PCM with low bit representation or frequency resolution sounds distorted to human ears as the square wave form of the discrete PCM values introduces perceptible harmonics. An encoding of 24 bit samples at 48 kilobits per second is standard practice for consumer audio signals. [4]

2.5 Amplification

The purpose of electrical audio signals is to accurately reproduce the recorded acoustic sound heard in a performance through conversion to electromechanical energy. This electromechanical record can be played back by loudspeakers which convert an electrical signal

back to acoustic waves through a magnetically driven speaker cone. Transmitted electrical audio signals must be amplified in order to be heard. Even in a signal chain whose audio information has been generated and transmitted in a purely digital form, there must at some point be a loudspeaker amplifying the digital signal for the end listener to hear the audio.

Electrical signal amplifiers are also essential to the audio recording process. Recording devices such as instrument pickups or microphones generate a weak signal. Any recording apparatus uses one or more amplifiers to increase the amplitude of a signal to something that is more useful in a signal transmission environment. However each stage of audio amplification introduces its own characteristic change to the signal.

Signals may be amplified by various electronic components. Usually one of three components is utilized to increase the power of an electrical signal: a transistor, a vacuum tube, or an opamp. Each of these technologies have their own electronics characteristic relationship between current and voltage. When a signal comes in whose amplification exceeds the capacity of the gain, the three components will begin to clip with three different characteristic patterns.

2.6 Distortion

When an amplified audio signal exceeds the threshold of the receptive input device, the waveform of the signal is dramatically altered. Any signal over the limit of the input is clipped to the maximum input value. The distortion of the waveform created by this clipping adds harmonics to the sound of the audio signal. Distortion is typically perceived as an undesired effect that should be removed from all parts of the signal chain. Analog electric guitar amplifiers however have been designed specifically to overload amplifying components so that they create harmonic distortion in certain musically pleasing ways. The design of electronic equipment

based on the characteristics of its distortion means selection of the amplifying component is essential.

2.7 Dynamics

The use of an electric guitar amplifier also introduces a natural level of compression to the signal output. This is especially the case for vacuum tube based designs, as the response to increasing gain is naturally compressed when processed through an overdriven tube. Compression is essentially a reduction in the dynamic range of a signal. The dynamics correspond to the energy or loudness of a sound. This corresponds to the amplitude of the signal.

A musical tone of any sort, whether voice, percussion, string, or anything else naturally exhibits a dynamic range which is termed the envelope. A distorted envelope or a synthetic sound without a dynamic range sounds unnatural to the human ear. The four essential stages of the envelope are attack, decay, sustain, and release. Attack corresponds to the increased excitement of the sound as it increases in energy. Decay is the immediate release of energy after the attack which then goes into a long hold of sustain before finally releasing. Each stage corresponds to a change in the dynamics.

Compression is a signal modification which reduces the dynamic range between these stages. This means that the quietest point of sound in the note is closer than the loudest point of sound in terms of amplitude. Compression in the electronics domain usually entails a reduction in the loudest parts of the sound relative to absolute quiet combined with an increase in the overall dynamics of all parts of the instrument. Compression was found to be useful for transmitting intelligible signals over telephone lines.

A good model of a vacuum tube will account for the natural compression that occurs when vacuum tubes are saturated with a strong signal.

CHAPTER 3: ELECTRIC GUITAR AND ELECTRIC GUITAR AMPLIFIERS

3.1 Stringed Instruments

Stringed instruments are a family of musical instruments which operate by exciting a taut string which resonates through a soundbox, creating a musical tone. Stringed instruments are among the most common of all musical instruments. The string can be excited by a plectrum, a hand, or through friction with a bow. The string creates a standing wave which defines the fundamental frequency and therefore the pitch of the tone being played. A musician may alter the pitch by placing pressure on the string against the neck of an instrument thereby shortening the length of the string being excited. The frequency of the string's vibration correlates to the fundamental frequency of the musical note to be played. A stringed instrument is designed to create similar tones for a string whose musical note value is being manipulated by the musician in order to create melody or harmony. The tone of the instrument is created by the shape of components of the instrument as well as the material of the string and the plucking device. [2]

3.2 Guitars

Guitars are a type of stringed instrument which are intended to be used with a plectrum or by hand, with a wood sounding box, and frets on the neck so that a player can more easily access the right tone for different notes. Guitars are designed with multiple strings, usually six, in the low-middle to middle range of tones. They are made to play chords, or multiple notes constituting a simultaneous harmony by having many strings pressed against frets in various shapes.

3.3 Electric Guitars

The electric guitar grew out of the jazz era of music where a guitar needed amplification to match the loudness of horns while playing together in big bands. Electric amplification was

seen as a way simply to make the sound of an acoustic guitar louder. As technology and musical tastes progressed, the electric guitar became a type of musical instrument separate from acoustic guitars and designed specifically to have a timbre shaped by the amplifier.

The essential operation of an electric guitar is to have strings made of metal. When the string is plucked, it excites an electromagnetic pickup situated on the body of the guitar. The pickup is typically placed near the bridge of the instrument where the hole of the sounding board on an acoustic guitar would be. A pickup consists of magnetic poles over each string, around which is a set of tightly wound metal wire. When the metal string of the player vibrates, the pickup produces a weak electrical current from the magnetic field of the strings. This electrical signal is routed to an amplifier which amplifies the signal. The amplified signal is used to drive a loudspeaker to make a sound.

The actions of the musician on the string create the fundamental frequency of a note as well as its dynamics. Although the body of the electric guitar will affect the timbre, the signal amplifier is a critical component in generating the timbre of the sound played. The range of fundamental frequencies of a standard electric guitar tuned to E ranges from 82Hz to 660Hz. The range of fundamental frequencies between 82Hz to 330Hz are most common. Some modern electric guitars tuned to lower frequencies may reach lower than 44Hz, though a guitar played at these frequencies is usually only observed for some specialized styles of playing in the Heavy Metal music genre.

3.4 Amplifiers and Distortion

Musicians and amplifier designers discovered a problem in amplifiers which turned out to be a defining feature of the sound of modern electric guitars. When an amplifier was turned up too loud, the circuit would not be able to handle the entire flow of electricity and the signal

would clip. This distorted the sound of the electric guitar. As mentioned earlier, clipping distorts a sound by adding non-linearities to the signal which add harmonics to the outgoing sound.

Amplifier designers began deliberately utilizing distortion to make more interesting sounds. A highly distorted electric guitar sound is characterized by metallic or aggressive sounding harmonics that are a consequence of the signal clipping and compression. A slightly distorted electric guitar tone is often given descriptive names like "adding crunch", "adding bite" or more. Modern electric guitar amplifiers also have controls for shaping the sound itself. Typically there is an equalization section with allows you to increase or decrease the availability of certain bands of frequency for any given note. There is also usually a gain control which gives the player control over the amount of distortion generated by the amplifier. [3]

The amplifier is such an important part of the modern guitar sound that different types of amplifiers are designed and marketed to different kinds of players. There is also a secondary market for other solid-state circuits that modify the signal coming into the amplifier, giving guitar players a wide array of available sound effects and timbres.

Although different kinds of amplification schemes exist, the second half of the 20th century saw the vacuum tube persist as the amplification component of choice for electric guitar amplifiers. Musicians and sound designers have not been able to replicate the sound of vacuum tube based distortion using solid state electronic components.

3.5 Digital Models

Digital modelling of vacuum tube based amplifiers has been a source of consumer interest because of the possibility of affordable quality tone. The first commercial digital modeling amplifiers were available to consumers in the late 1990s. Most digital modelling research of amplifiers is driven by commercial interests and not available to the public.

Generally, the models convert the signal from the electric guitar directly to digital, then use digital filters or waveshaping functions on the input in order to add distortion and harmonics to the original sound. This digital emulation of an amplifier can either be used natively in a digital recording environment or sent to a solid-state power amplifying unit directly to a loudspeaker.

Reception of digital models for guitar amplifiers has been mixed. Most agree that they do not sound precisely like vacuum tube based amplifiers, nor do they respond to the performance of a musician with respect to dynamics and tone the same way. This is because of the non-linear distortion characteristics that are impossible to completely replicate in an analog amplifier. All digital models suffer as well from undesired digital distortion generated by aliasing. Aliasing is the inability to capture the tone completely due to the limited sampling rate of a digital model.

The experience of digital models and their relative lack of realism leads to an interest in machine learning as a solution. This fits within the pattern of other tasks which a machine learning approach succeeds. The amplified output of vacuum tubes introduces many non-linearities to the system. Machine learning algorithms are proficient at understanding and repeating non-linear patterns.

CHAPTER 4: BASICS OF MACHINE LEARNING

4.1 Definition of machine learning

Machine Learning is a subset of artificial intelligence which concerns itself with training a computer to learn a program instead of trying to manually program logic. Machine learning has taken on immense interest for countless breakthroughs in accomplishing tasks that are difficult to program on a machine yet seem intuitively simple for humans.

A general, commonly accepted definition of Machine Learning from Mitchell is that "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks T, improves with experience E." Machine learning uses statistical concepts to describe a task as well as the performance of a solution. [1]

4.2 Machine learnings tasks

A task in machine learning may be described as belonging to one of a class of problems. One class is discriminative, which means selecting a classification for some particular piece of data. Another class of problems is generative, which means generating an output that might fit into a particular class, whether or not the class is defined. [1]

The task is usually defined by what we want the machine learning program to do with examples. Examples, that is the underlying data we learn from, are represented in a vector of values of the same dimension as other examples. Example data are often called features.

The purpose of a machine learning task is to discriminate between classes from input or to generate output so that the correct classification or generational attributes can improve statistically by learning from training input.

4.3 Performance of the task

A machine learning program is evaluated according to its performance in how we want the task features to be processed. For example, if we want the machine learning program to classify a set of points in Euclidean space between two classes, we would evaluate the performance according to a percentage of how many features were correctly assigned to their actual class. [1]

4.4 Supervised and unsupervised learning

There are two general classes of machine learning tasks, supervised and unsupervised. These two classes are differentiated by whether or not the example features from which a program learns are labeled with a class of solutions. If we only have a set of features and no underlying classification, then we cannot teach a program in any way that we could measure the performance of a task. Since machine learning is based on statistical distributions, we could attempt to learn underlying statistical distributions of feature sets and then measure the performance on a test set of features if the learned distribution fits the new test data. [1]

A supervised machine learning task is one where an algorithm learns a function for mapping an input to an output based on previous examples of inputs and their corresponding outputs. This information is called labeled training data, as the inputs have been labeled with the corresponding output they represent. An unsupervised learning algorithm is one where the output labels to a given set of inputs are not given, so the algorithm must learn a set of classifications for mapping inputs to a discovered set of outputs.

4.5 Evaluating the model

In order for a machine learning algorithm to be effective it must be able to generalize its learned parameters to features beyond the set it learned. Generalizability is different than the performance of a task because it is possible to train a model in such a way that it could have a

low error rate but perform poorly on new features even if those features are actually random variables from the same underlying probability distribution as the learned features. [1]

This means there are actually two different kinds of error in training a machine learning model. The first mentioned earlier is the training error. The training error is the error of the machine learning model not optimizing completely on the training set. The training error must be reduced in a way that optimizes. [1]

The second type of error is generalization error. When a model is tested against data for which it has not trained we can generate an error rate for its performance. It is possible for a model with lower training error to perform much worse than a model that is not as optimized when testing against a new data set. This means the machine learning program has learned a statistical distribution that is in some way incorrect. Underfitting occurs when we have not properly optimized the training error of the model. Overfitting occurs when the model is optimized but the test error is still high. [1]

The goal of designing a machine learning program is to design a learning algorithm that can perform well on a specific task. This means we need to know as much about the underlying data as well as the nature and assumptions of a task to make judgments on what approach we will use to find optimal results.

4.6 Linear Regression with a loss function

The simplest example of machine learning is to train a linear regression. In this task we are given a list of vectors in euclidean space. For simplicity in the example we will explain in two dimensions. The goal is to learn a function that best predicts for any vector (x,y) the coordinate y given its coordinate x. For a linear model we will want to learn the parameters a, b of the function $y = ax + b$.

In order to define what we want to learn, we must define what is a good fit. This is accomplished through a loss function. For the linear regression case we can define our loss as mean squared error (MSE), a statistical measurement that squares the errors of all points from their predicted values of our model, and finds the average of these. The best fit we can find is one that minimizes the loss. We modify our a and b values until we find the minimum on our loss function. [1]

CHAPTER 5: NEURAL NETWORKS

5.1 Purpose of feedforward neural network

A feedforward neural network is a particular type of machine learning model that, like other machine learning models, tries to approximate a function. The feedforward neural network approximates a function y=f(x, theta) where x is the input of the function and theta are the set of parameters that best approximate the function. Neural networks have been found to work very well for non-linear patterns in data. This accuracy in non-linear patterns has made neural networks particularly well suited to classification and generation tasks related to human perception. As musical tone is a matter of human perception, this project uses a neural network architecture to train machine learning models. [1]

The name feedforward neural network refers to two particular features of the model. First, the model is a combination of function compositions. Each function and its relation in the composition of the approximated model can be visualized as a graph. These compositions form layers, where a layer of functions is connected to all of the others in the next layer, up until the output layer provides a result. There is no feedback mechanism among the compositions, so the network is considered feedforward. In formal algorithmic terms the function compositions and their dependencies create a directed acyclic graph. Second, the relationship between these functions is modeled on the activation behavior of biological neuron cells. Each function describes a relationship to its input activations, which then creates an output to an activation function. An activation function is a special kind of function, usually non-linear, that gives a value of between 0.0 and 1.0.

5.2 Perceptrons Activation Function

A neural network is essentially a composition of functions. The functions interact with each other by way of an activation on functions that one function is connected to. The output of a particular neuron in the network is defined by its input and the function definition. If this output meets the criteria of activation, then the activation output is sent to all other connected inputs. With this structure a neural network can learn a complex relationship between the input and output training data.

5.3 Graph model

The connections formed by the complex compositions of learned functions can be seen as a set of nodes connected in a graph by one-way edges. The nature of the feedforward design means the graph model of a neural network is a directed acyclic graph.

5.4 Output, Input, and Hidden Layers

The nature of the problem which a neural network solves lends itself to organizing the so-called neurons of the neural network into layers. The input layer represents a set of activations of the particular input being evaluated according to the structure of the elements of the input vector. For example, 2x2 pixel image in black and white may be represented by 4 input activations which may represent either white or black.

The output layer of the neural network must represent the structure of the output desired of the solutions. For example, a network trained to categorize an input into any of four categories would have four outputs. Every layer in between the input and output layers where there is a mesh of connectivity between input nodes and the nodes of the layer represents a hidden layer. A complex model may have many hidden layers of varying sizes or connectivity.

5.5 Training a Neural Network with Backpropagation

A neural network can be trained by evaluating the values of an input vector to whether the model accurately predicts the correct output and adjusting to minimize the loss on a set of inputs. As mentioned earlier, the nodes represent one element of a complex composition. We can adjust the loss function according to a gradient on each node by propagating the calculation of loss backwards along the directed acyclic graph structure of the model.

CHAPTER 6: RECURRENT NEURAL NETWORKS

6.1 Definition and purpose

As mentioned earlier a typical neural network has a vector set for the input and output layers. This will work well for inputs arranged in finite euclidean space. For a time series or other signal type input and output the structure of the standard neural network will not fit. The solution is to provide a structure of neural network whose weights for each node are identical for each time step and then provide a way for the graph structure at a previous iteration to feed into the graph structure of the current time as an input. A recurrent neural network is a neural network whose graph model provides feedback. [1]

6.2 Classes of Recurrent Neural Networks

There are three classes of recurrent neural networks. The three classes correspond to whether the input or output is a vector of a single value or is a stream of recurring values. This lends itself to a one-to-many, a many-to-one, or many-to-many transformation. One-to-many may be thought of as generation of a signal based on an input where a single vector input leads to the recurrent neural network structure to generate a series of time valued steps. Many-to-one may be regarded as signal categorization into finite categories. Many-to-many may be thought of as a signal transformation function where the neural network evaluates each step of a series and predicts the correct output of the transformation in series as well. [1]

6.3 Forward propagation equations

The recurrent neural network structure may be trained on corresponding input and output training data the same as a regular neural network through backpropagation of errors. This is because all recurrences of previous times appear in the neural network graph as edges and retain the same directed acyclical properties of the regular layers of neural networks.

6.4 Long term dependencies

The main concern of research into recurrent neural networks is how to capture relations between recurrences of the neural network that are far apart in time. The farther apart in time two nodes are, the more difficult it is to capture the relation between the signals even if there is actually a very strong correlation present in the training data.

6.5 The WaveNet Model

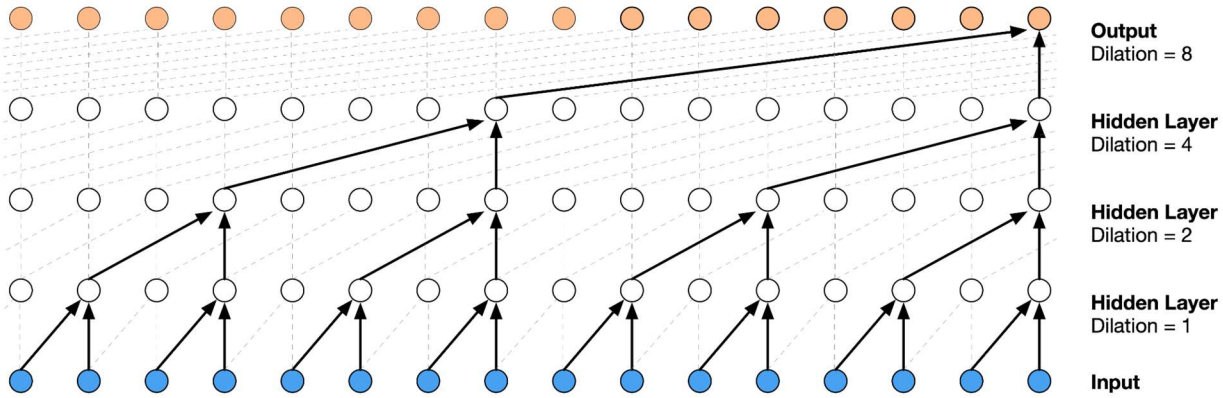


Figure 1: Visualization of WaveNet model from VanDenOord et. al. [5]

Recent research for acoustic signals, particularly the human voice, has found that capturing long term dependencies in a series representing pulse-code modulated digital audio can be done with a convolutional layer model. A convolutional layer only takes as input a subset of the previous layer. The WaveNet model uses a convolutional architecture based on each hidden layer representing a dilation of the output to more and more inputs over time. [5]

CHAPTER 7: EXPERIMENT DESIGN

7.1 Emulating an analog amplifier and recording chain

For purposes of recording an electric guitar as played through an amplifier, the effect of an amplifier on a guitar signal can be seen as a transformation function from an input signal to a final signal. A traditional studio arrangement is to have the musician plug into an electric guitar amplifier whose speaker cabinet has a microphone in proximity. This recorded signal is then mixed into a final track for listeners. Modern digital technology allows the unchanged dry signal coming from an electric guitar line to be fed directly into a digital system. Software plugins are able to transform the signal in a similar manner as the entire signal chain. A modified WaveNet model has been shown to successfully replicate the effects of components on an electric guitar signal. [6] Researchers have devised a "PedalNet" model meant to adapt the WaveNet model to the domain of musical instruments. Pedal Net differs from WaveNet in that it does not contain a softmax bin of 256 outputs and instead models directly the amplitude of each input point. [6] PedalNet also accepts as hyperparameters the dilation size and the number of layers present in the convolutional WaveNet model.

This experiment aims to explore the findings of previous literature on "PedalNet" by confirming, expanding, and optimizing. We aim to confirm the findings of the previous literature that a convolutional recurrent neural network is effective as an electric guitar amplifier emulation. We aim to expand upon the previous literature by introducing more complex and rich harmonics into the amplifier signal chain to see if the PedalNet architecture can successfully capture denser harmonic distortion signatures than previously tested. We also aim to optimize by first, finding the best practical design of inputs and outputs to train a PedalNet for optimal results and second, by suggesting an alternative architecture for PedalNet that could capture denser

harmonics. The experiment intends to test the models on a complete studio recording setup and not just individual components in isolation. Finally, subjective listener tests are performed to evaluate the effectiveness of the digital model signals in a fully mixed audio recording with multiple instruments present.

7.2 Signal chain setup

For this experiment a dry guitar signal was captured as an input to training data for the model. The same signal was then captured after it traveled through a complete amplifier chain of a booster pedal, a 100W vacuum tube based electric guitar amplifier, a high volume setting to fully saturate the power section of the amplifier, an attenuated cable connection to one 12 inch speaker to control volume, a microphone, a microphone preamp, and a compressor unit. This processed signal was used as the output training data. All inputs and outputs were recorded at 32bit 44.1 kHz sampling rate. All models were trained with combined recordings between 233 and 239 seconds. This is the same amount of training data recommended in the previous literature. [6] A reamplification unit was used to be able to record an input and then play the identical input to different output settings.

Signal Chain of Trained Models

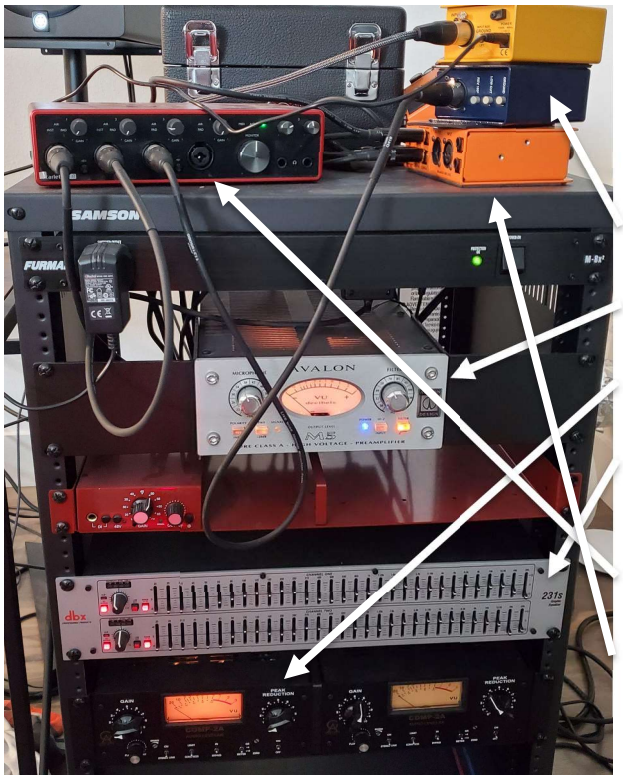


Figure 2.1 Left: rackmount of studio recording chain used to record training output signal and experiment control tracks

Rackmount Component and Description

1. Blue box, Direct Box, used to record dry guitar signal while playing
2. Gray rackmount, Preamp, used to amplify incoming microphone signal
3. Black rackmount, Compressor, used to smooth signal transients from preamp
4. Gray large rackmount with band controls, EQ used to lower volume from compressor to avoid clipping
5. Red box, digital interface, collects signal from EQ unit and digitizes to send to computer
6. Yellow box, 'Reamp' Box, used to send pre-recorded dry signal to amplifier chain

Figure 2.2 Left: amplifier used to record training and experiment control tracks

Amplifier Component and Description

1. Orange Rockerverb 100 MkIII, Amplifier
2. Purple box, load box, set to 16 ohm to absorb half the power of the signal from the amplifier
3. Speaker cabinet, 1x12 speaker 16 ohm load plugged into the amp output in parallel with the load box
4. Amplifier microphone sent to rackmount preamp

Figure 2.3

Pre-amplifier pedal settings

1. Chorus effect used in Jazz setup for wider sound and gentle modulation
2. hard-clipping overdrive used in Blues setup with gain set to provide slight breakup before the amplifier
3. 'Tubescreamer' type overdrive used in Rock setup to tighten the low and high end as well as amplify the incoming signal as much as possible

Figure 2.4

Figure 2.5

On left:

Figure 2.3: Stereo polychorus pedal, used before amplifier in Jazz track to provide width and warmth

Figure 2.4: EHX GLOVE overdrive, used before amplifier in Blues track for standard hard-clipping 'blues style' overdrive to add grit to a clean.

Figure 2.5: Maxon ST9 Pro+ overdrive, used before amplifier in Rock track to tighten the sound and to add power to push the amplifier preamp section into maximum distortion

On left: Figure 2.6: 'Dirty' channel pre-amp settings used for Rock track. Volume (first knob) on high to saturate the amplifier power stage. Pre-amp gain (fifth knob) set to breakup

Figure 2.7: 'Clean' channel pre-amp settings used for Blues and Jazz tracks.
Volume on high to saturate power stage

Figure 2.8: Power stage settings for all tracks
Reverberations off. Attenuator set to high to lower overall loudness while recording

7.3 Musical categorization of collected inputs and outputs

Three amplifier settings and their corresponding dry guitar inputs were selected. Two input training data tracks were used. The first used only electric guitars that featured 'single-coil' pickups. Two guitars were used with single coil pickups in the bridge as well as the neck position with a variety of sounds. The training input file for this category was 233 seconds long. The second input training track used only electronically active double-coil pickups in the bridge position. These pickups utilized ceramic magnet cores for an aggressive attack and loud presence. The second input training track used guitars with a variety of lowered strings and drop tunings. These guitar configurations are more commonly used in so-called hard genres such as rock and metal. The training input file for this category was 239 seconds long.

The three amplifier configurations were chosen based on typical settings for a particular genre of music. The three genres were blues, jazz, and rock. Blues used the clean channel of the amplifier which did not utilize any preamplifier distortion. The blues model used a hard-clipping overdrive pedal connected to the input of the amplifier after the guitar for a soft distortion. This configuration is often used in blues as it offers some 'grit' in terms of distortion, but the grit is softened by the warmth of the sound provided by the vacuum tube amplification of the amplifier. The goal of a blues style tone is to be smooth and full but to have some harmonics added to the guitar so that it can stand out as a lead tone. The blues model used the single coil input data.

The second model was designed for jazz. This model used the single coil training input as well as the

Guitars used for Jazz and Blues models

Figure 3.1: Guitars used to record training and testing tracks for Jazz and Blues models. Only 'single-coil' pickup style was used. The pickups on all three guitars are different models.

Training	Play samples and Testing
	

Figure 3.2: Guitars used to record training and testing tracks for Rock model. All guitars feature 'Fishman Fluence Modern' Active pickups with ceramic cores for high output from guitar.

Training	Play Samples and testing
	

7.4 Training and refinement

The models were trained on high performance computing systems provided by the University. Most models took between two to four hours to train using one graphics processing unit. For each successive model, a track was developed using a guitar which was similar in specifications compared to the guitars used for the relative training data. For the jazz and blues models, a guitar with a single coil neck pickup was used which featured a different kind of electronics configuration than the training guitars. For the rock model, the testing used an 8-string guitar which had an identical bridge pickup model as the training guitars but with different body construction and different string tuning. The model outputs were played back in a studio monitor setting for continuous refinement. Top performance required listening analysis of many different model and input settings.

7.5 Development and preparation of training data

1. Use of similar instruments

It was found necessary to limit the variety of input types in the input training data for best results. In practice this limitation makes sense as guitars designed for a certain style of play are typically paired with complimentary amplifier settings for the desired tone characteristics. The criteria of using similar instruments in the input training data was discovered as necessary after observance of poor performance when too many different types of guitars were used as input. The previous literature used a set of standardized testing that only featured two guitars. The inputs did not feature any type of playing techniques other than open strumming of single notes or chords at the same dynamic level. This led to predictable results for the outputs.

For this experiment there were a much more varied group of guitars available as well as recording of different techniques such as palm mutes, pinch harmonics, string harmonics, and

pitch bending. The different types of guitars, and particularly the relative strength of the signal coming from the pick-up electronics provide very different input signals. All of these signals were compressed in dynamics and distorted in a similar way. This created a situation where many different input types led to underfitting of output as the model could not create the varying dynamics of the electronics as well as the amplifier at the same time. Limiting the input to similar guitars, particularly the output power of the pick-up electronics significantly improved performance.

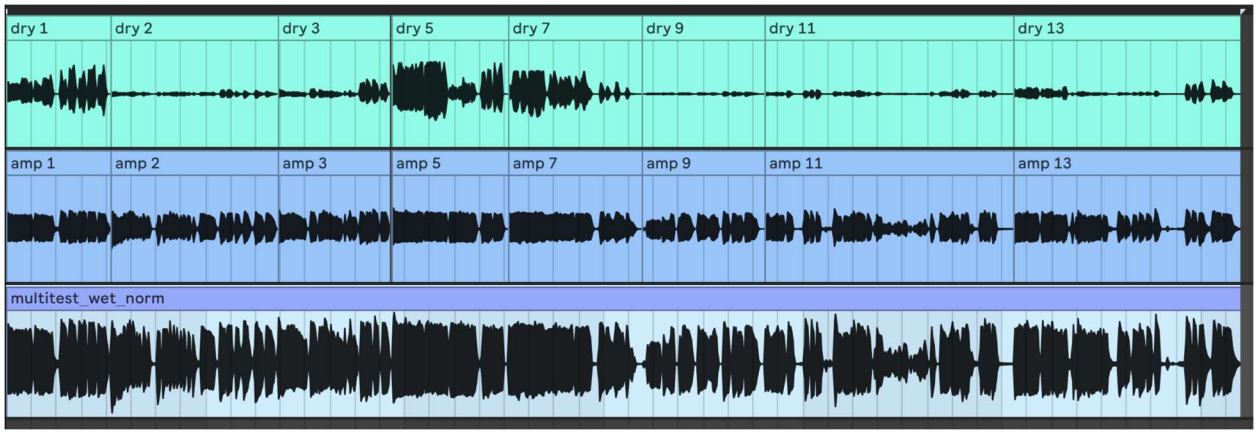


Figure 4: Input training data based on 8 different guitars and two outputs from different amplifier settings in the same recording signal chain. Compression of both weak and strong input signals to equivalent output creates a difficult pattern for the WaveNet architecture to model.

2. Phase correction

It was also discovered that the convolutional recurrent neural network model benefited substantially in audio quality when the output training data was phase corrected to the input. Phase correction means manually examining the digital data of two corresponding tracks and adjusting the phase so that the phase of the two signals is in alignment. In practice this meant moving the output wave back in time relative to the input. The action of the complex components on the output signal meant the output came in at a fractionally different time. Even a

very small phase difference could mean the output was off by 400 or more samples from the input.

For the purposes of a convolutional neural network, missing important information corresponding to the beginning of the instrument's sound pulse becomes a problem because the receptive field is shortened, and the most sensitive area of the model's pattern learning cannot be utilized. The receptive field is shortened because the model learns that the only relevant pattern for the output is what is happening not at t-1 but at a sample farther back in time.

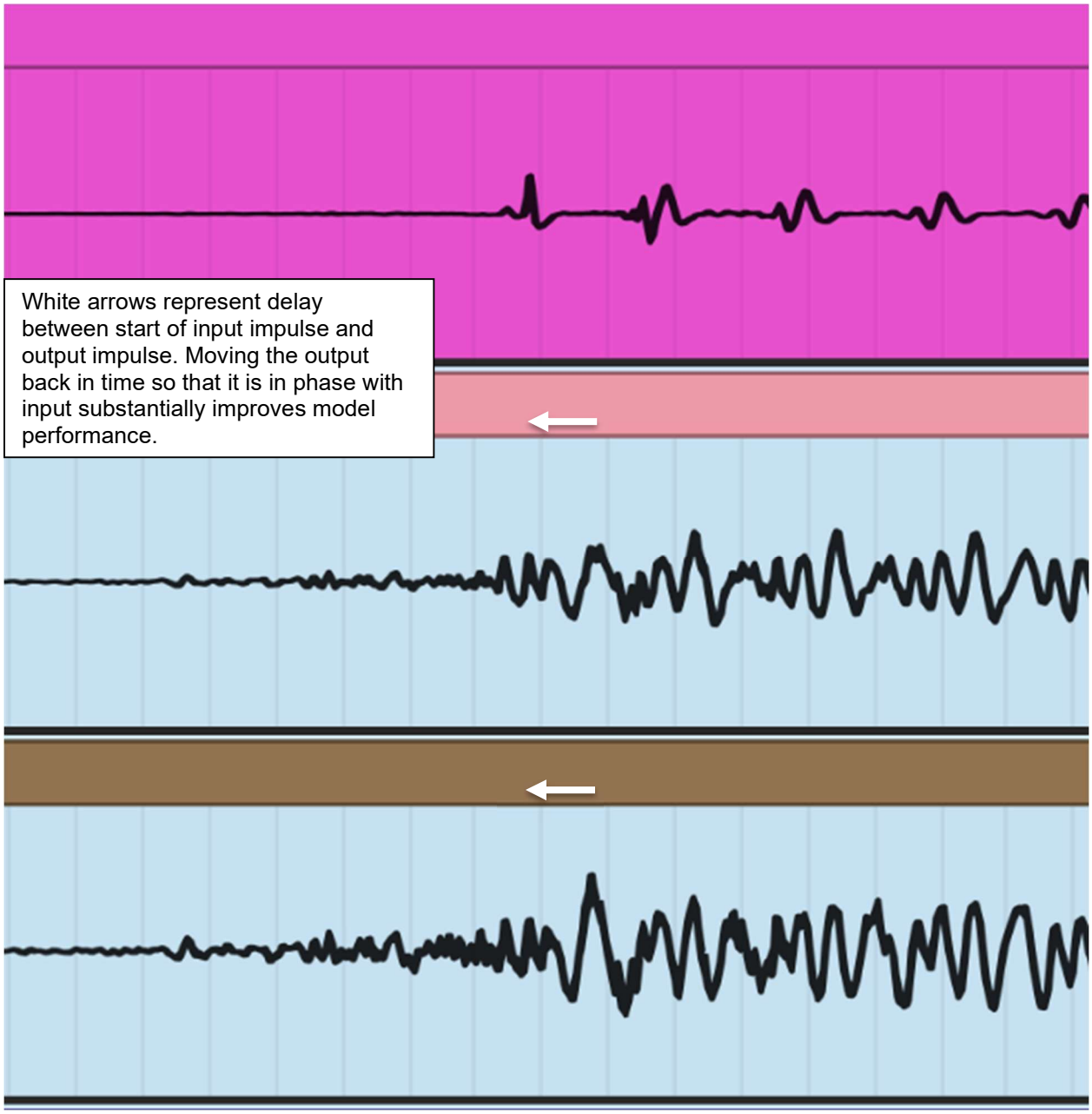


Figure 5: Input and two output models before phase correction. The time gap between the input and output (see signal peaks for reference) led to observations from audio monitoring of a glitching sound during fast playing. Phase correction before training a model on the inputs and outputs proved very beneficial to overall sound quality.

7.6 Hyperparameter settings for test models

Three different hyperparameter settings were used for each of the three different models. The three hyperparameter configurations were labeled as minimum, mid, and maximum. Only the minimum model was able in any model to be played musically in real time on a digital audio workstation. The mid and maximum models took extraordinary times to load and were essentially used as post-processing steps done in the studio. The minimum model used 12 convolutional layers with a dilation of 8. The mid model used 32 convolutional layers as well as a dilation of 8. The maximum model used a total of 64 convolutional layers with a dilation as well of 8. The dilation size was discovered to have an adverse effect on model quality for mid and maximum as it was extended beyond 8. Increasing the dilation size on the minimum model made it not perform in real-time.

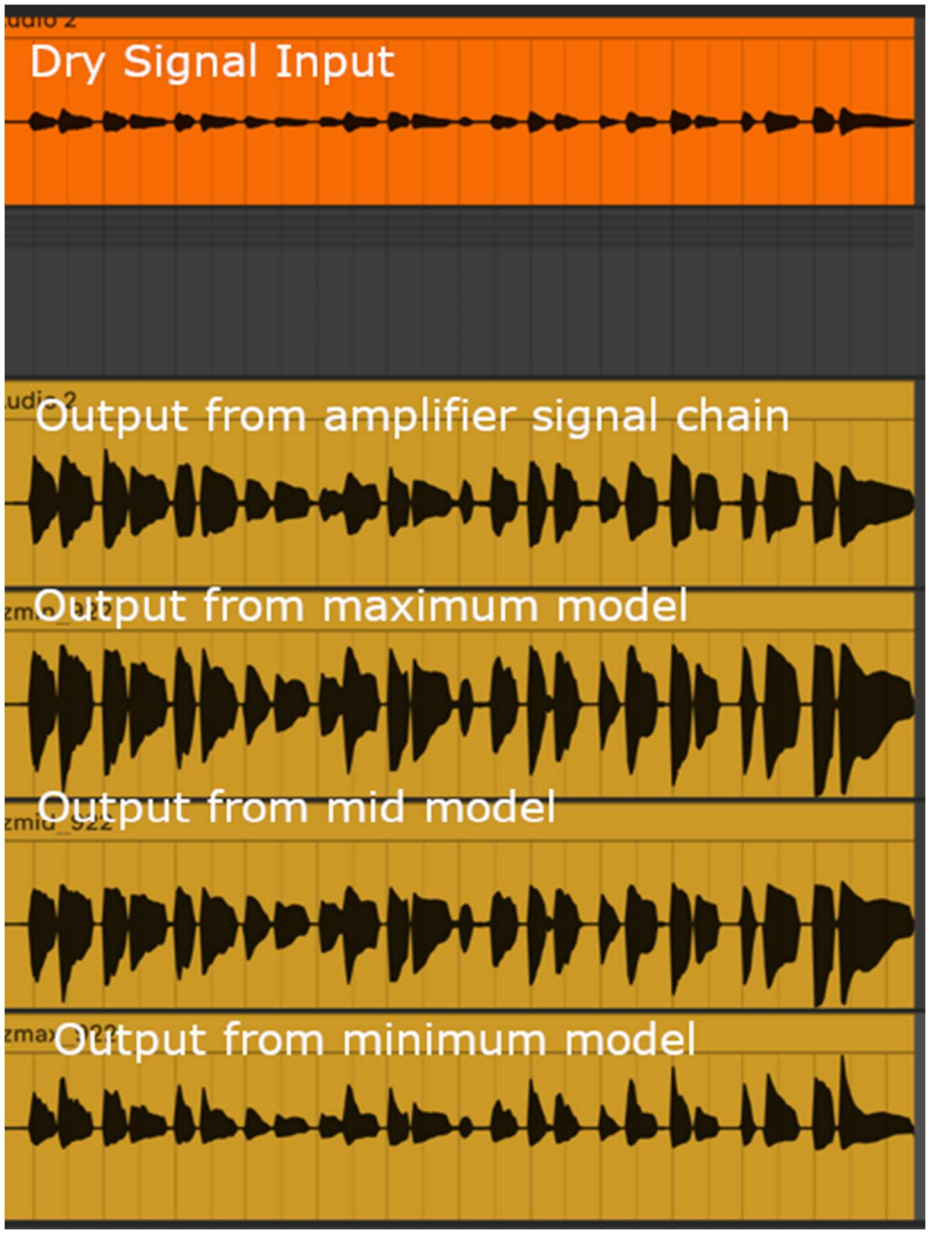


Figure 6: View of guitar tracks inside session. Each track was mixed and mastered with identical bass and drum tracks to allow for comparison within a finished musical audio track.

CHAPTER 8: TESTING AND ANALYSIS

8.1 Signal Analysis

Signals were analyzed by looking at spectral representations of each amplifier and model output to compare. The regular spectrogram shows a general view of the harmonic characteristics of a sound. The melodic range and peak frequency spectrographs are designed to show the most prevalent harmonic patterns and correlate them to specific musical intervals from the fundamental frequency of the note being played. The output analysis was from testing input from playing a guitar which was not a part of the testing data. The most obvious patterns to show up by the spectrogram analysis were the lack of definition in high frequency harmonics in all models, not capturing the 'ripple' effect of distorting vacuum tubes in high gain models, as well as a less dynamic response to playing.

	'Blues' Model: Clean Channel, hard-clipped Overdrive, single-coil pickup

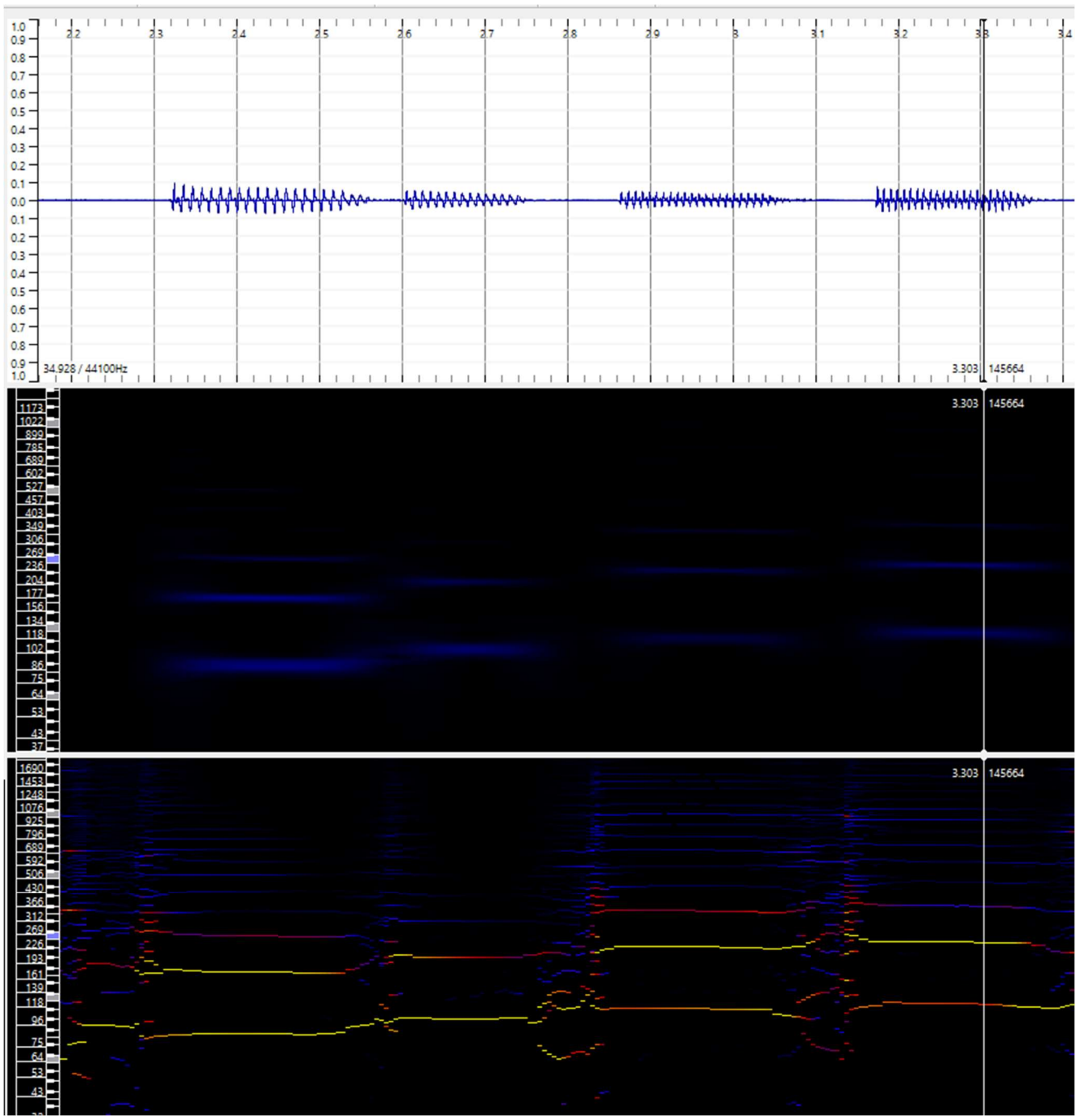
Table 1: Dry Guitar	 dry_blues.wav

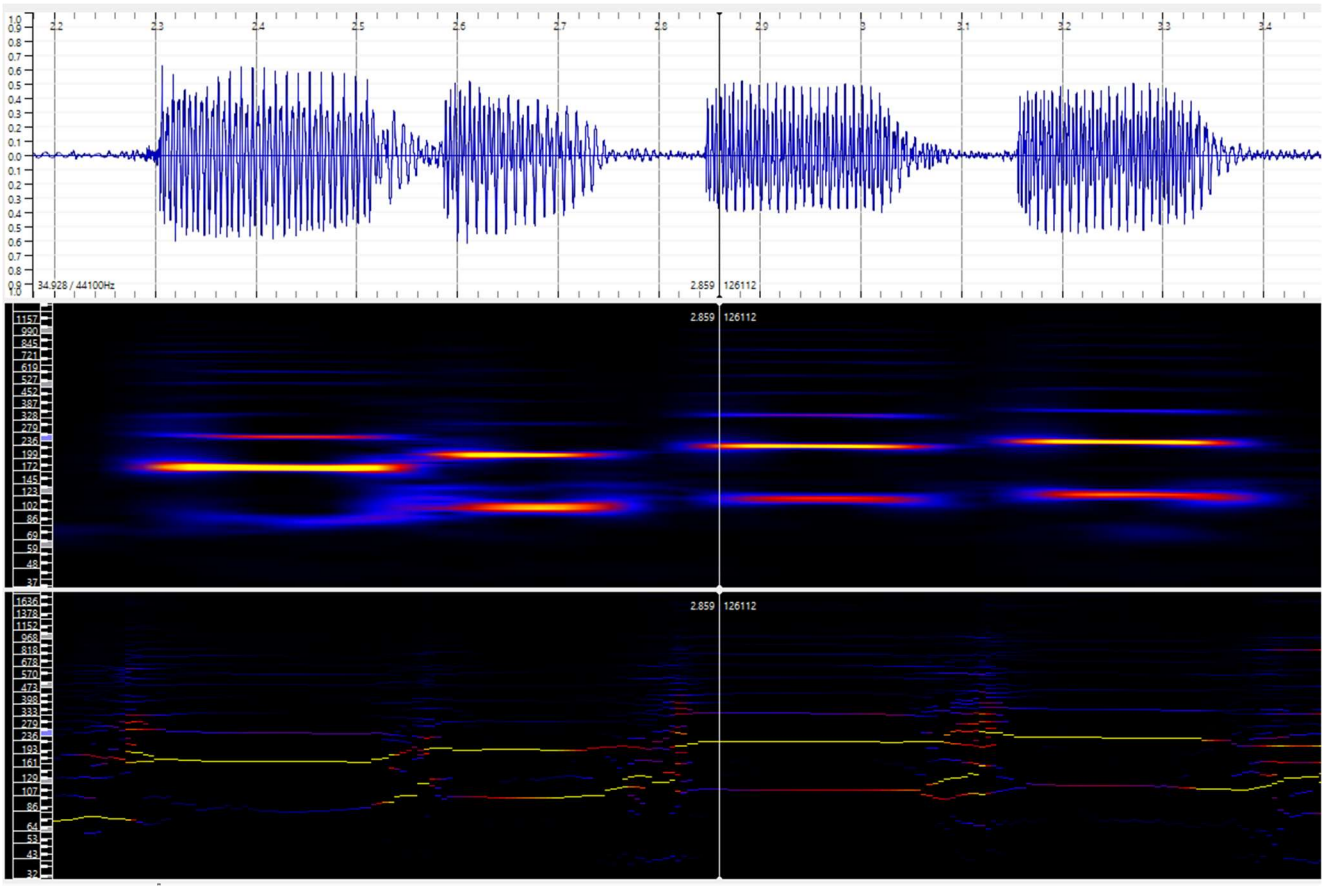
Table 2: Blues Model Amplifier	 soundtest_blues_amp.wav

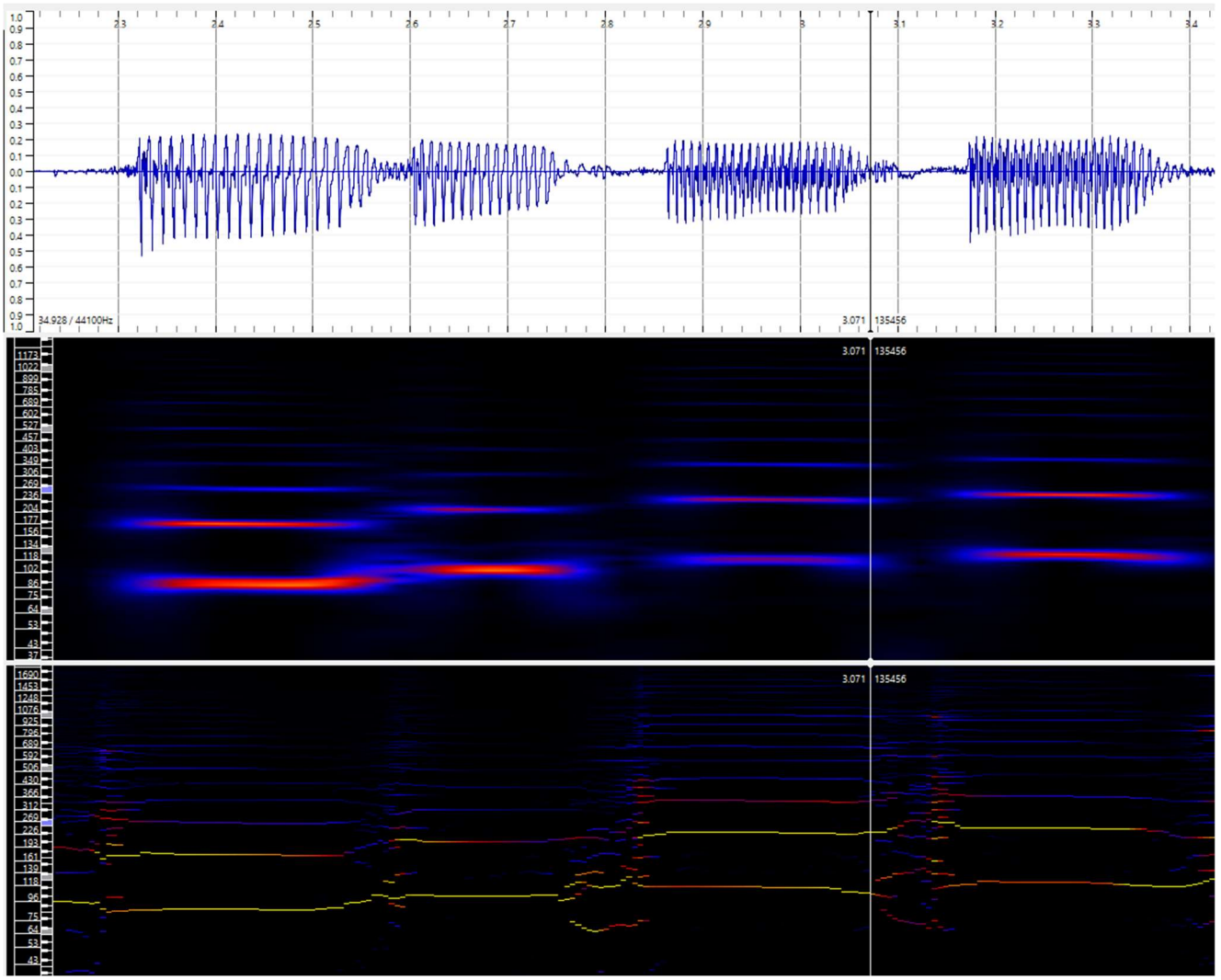
Table 3: Blues Min Model, 12 layers	 soundtest_blues_min_model.wav

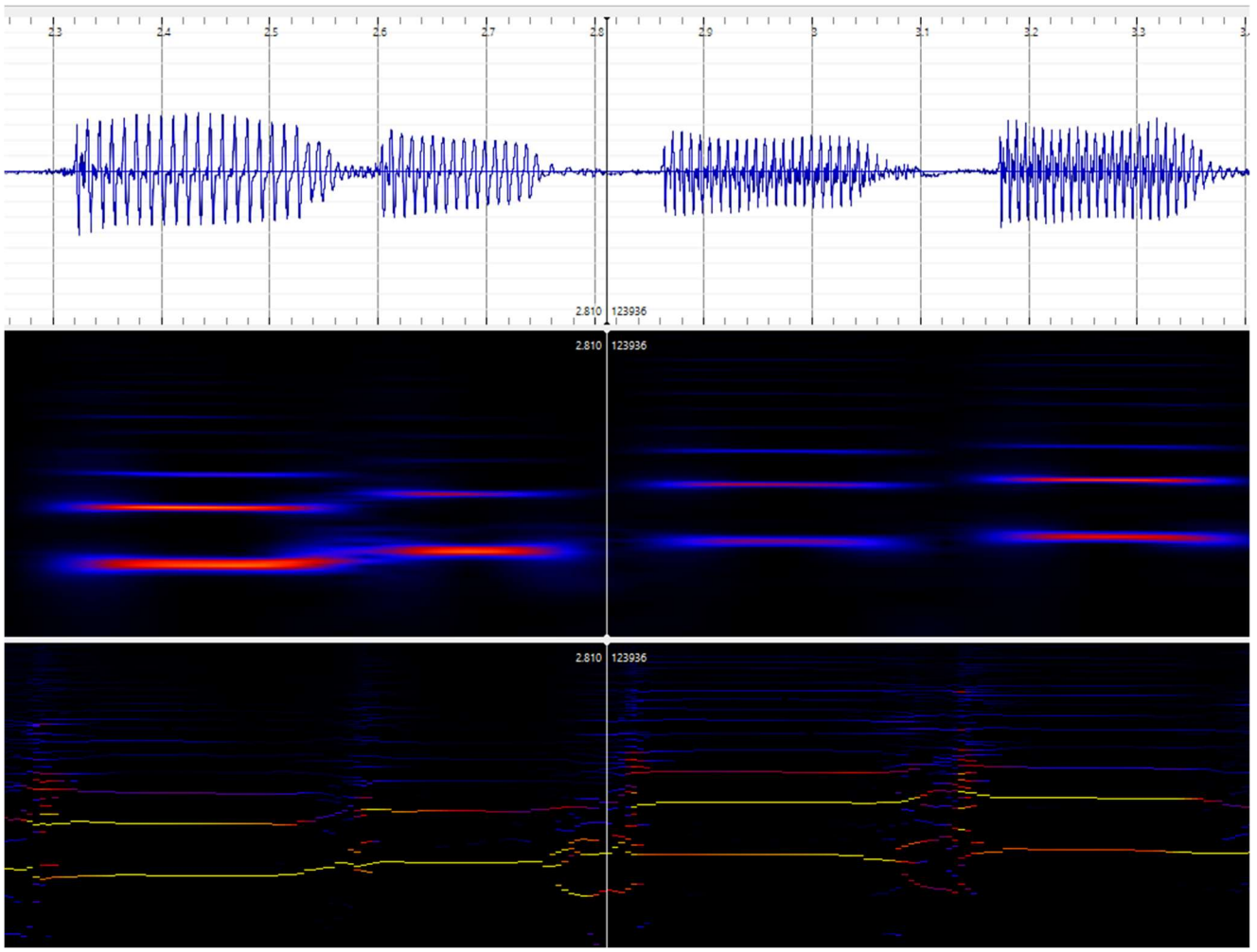
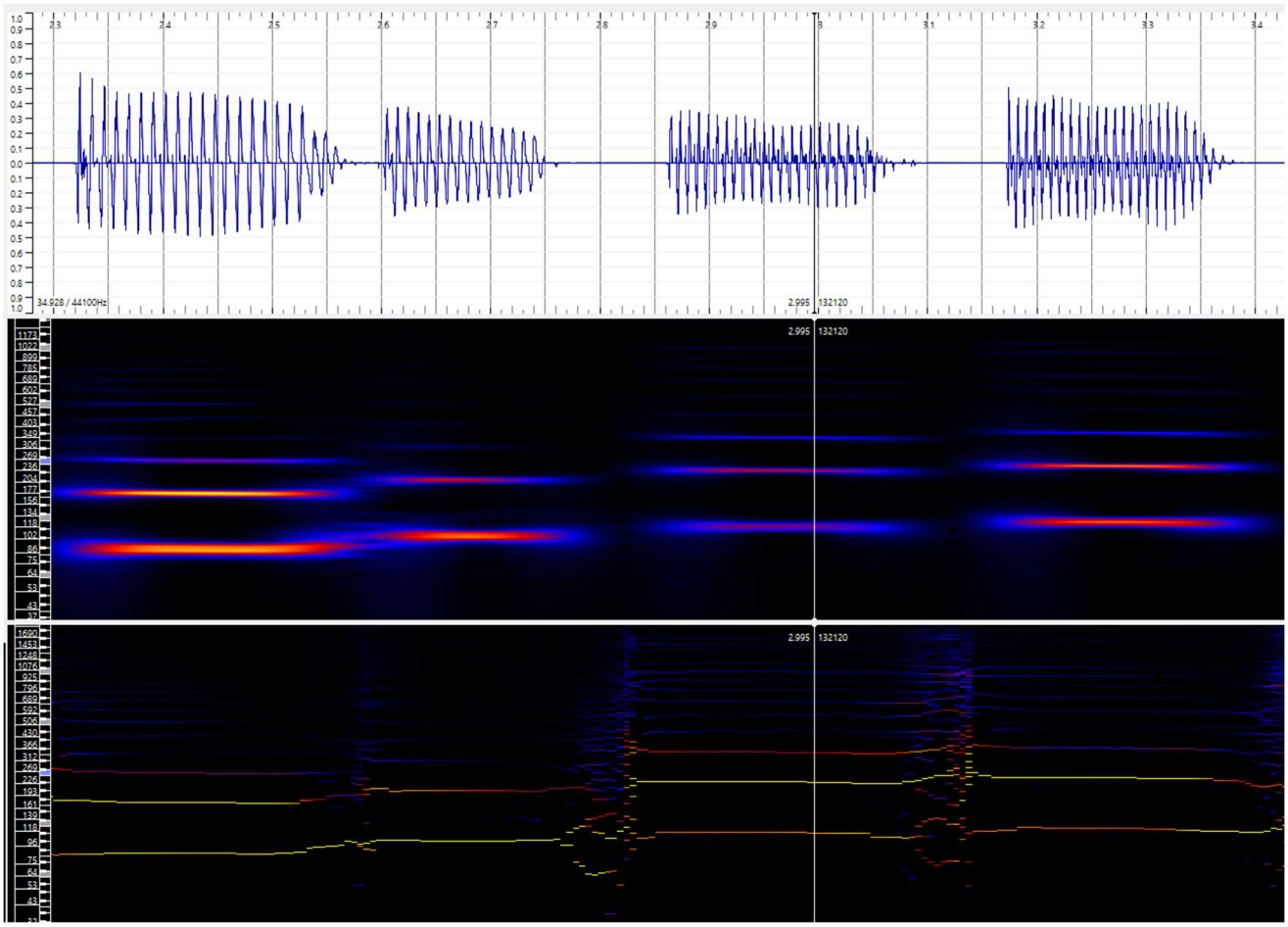
Table 4: Blues Mid Model, 32 layers	 soundtest_blues_mid_model.wav

Table 5: Blues Max Model, 64 layers	 soundtest_blues_max_model.wav

	'Jazz' Model: Clean Channel, chorus pedal, single-coil pickup

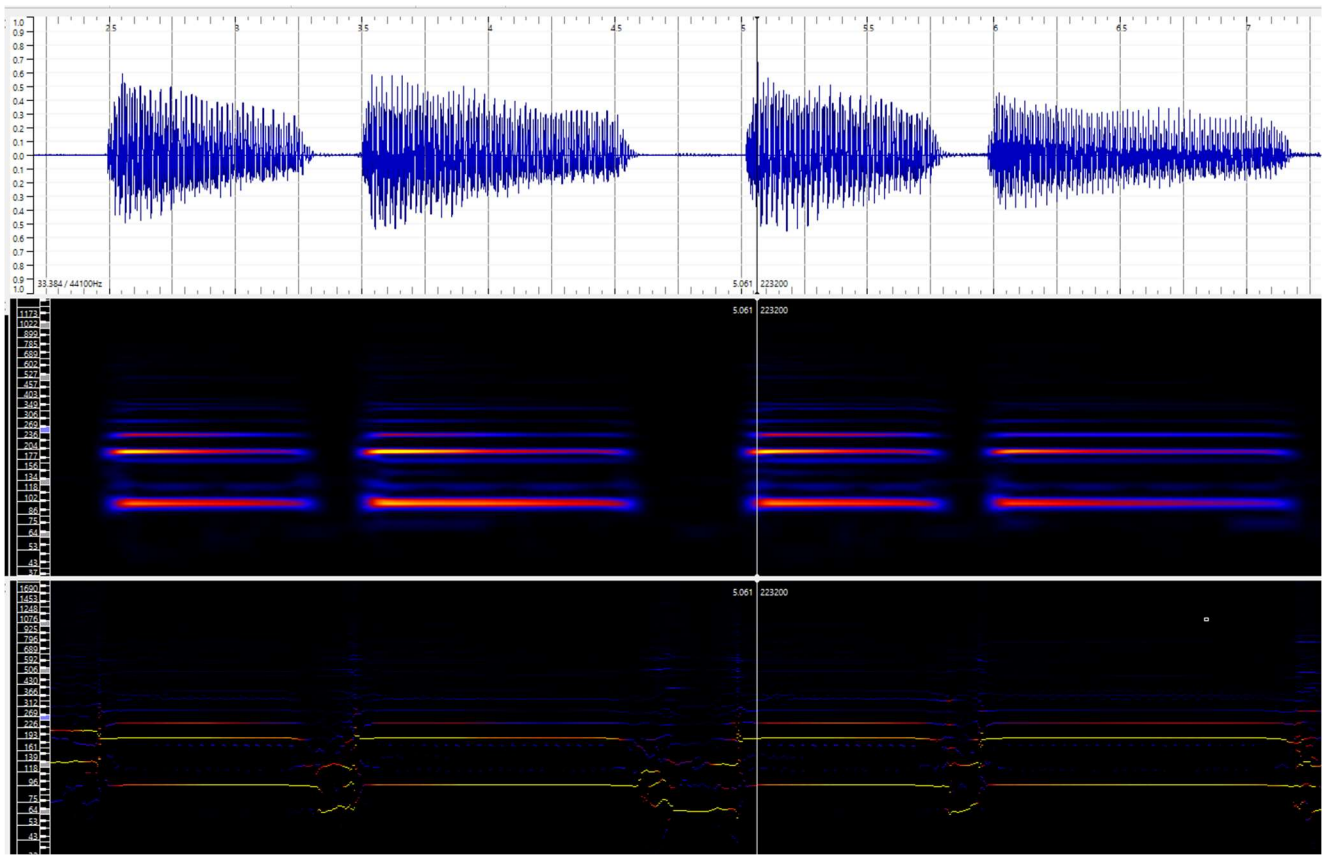
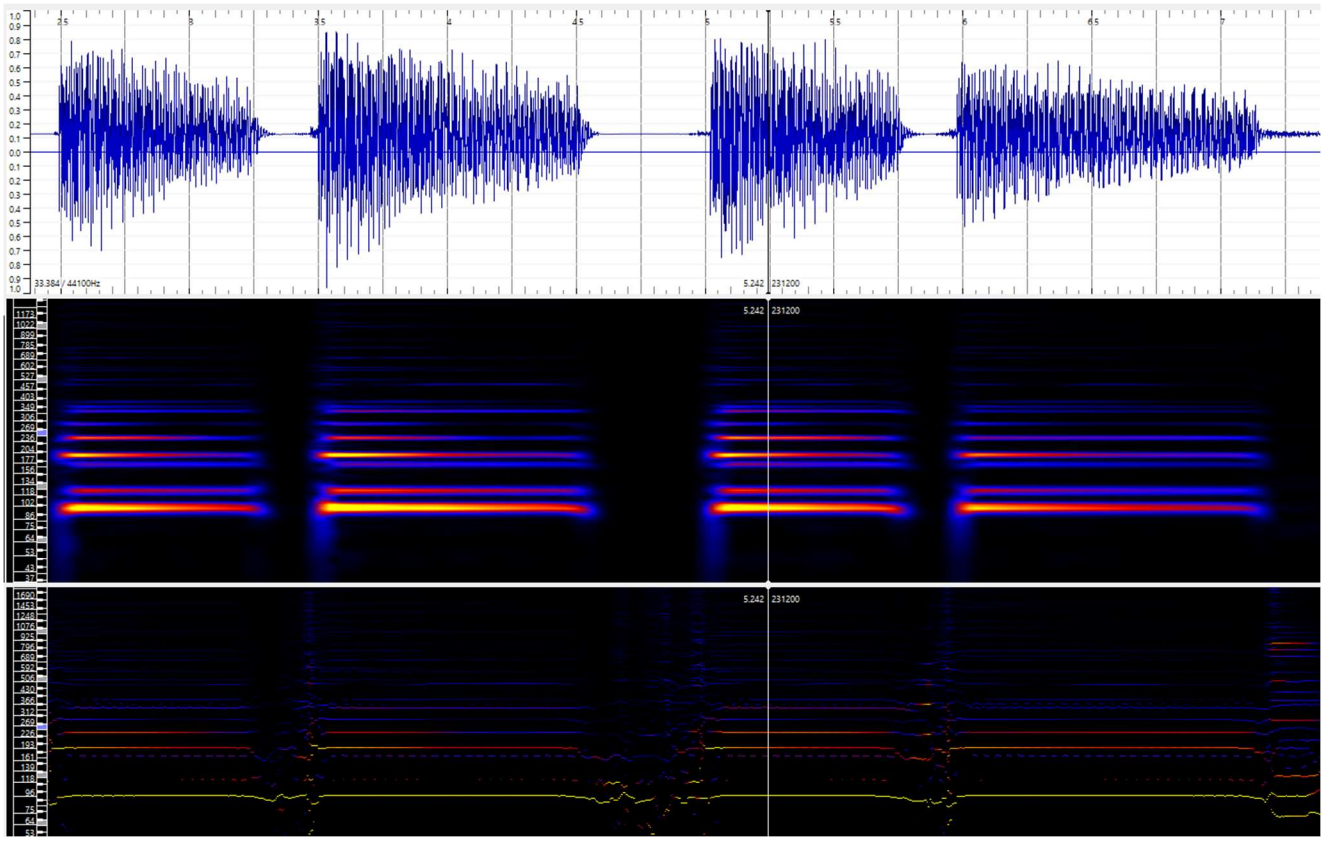
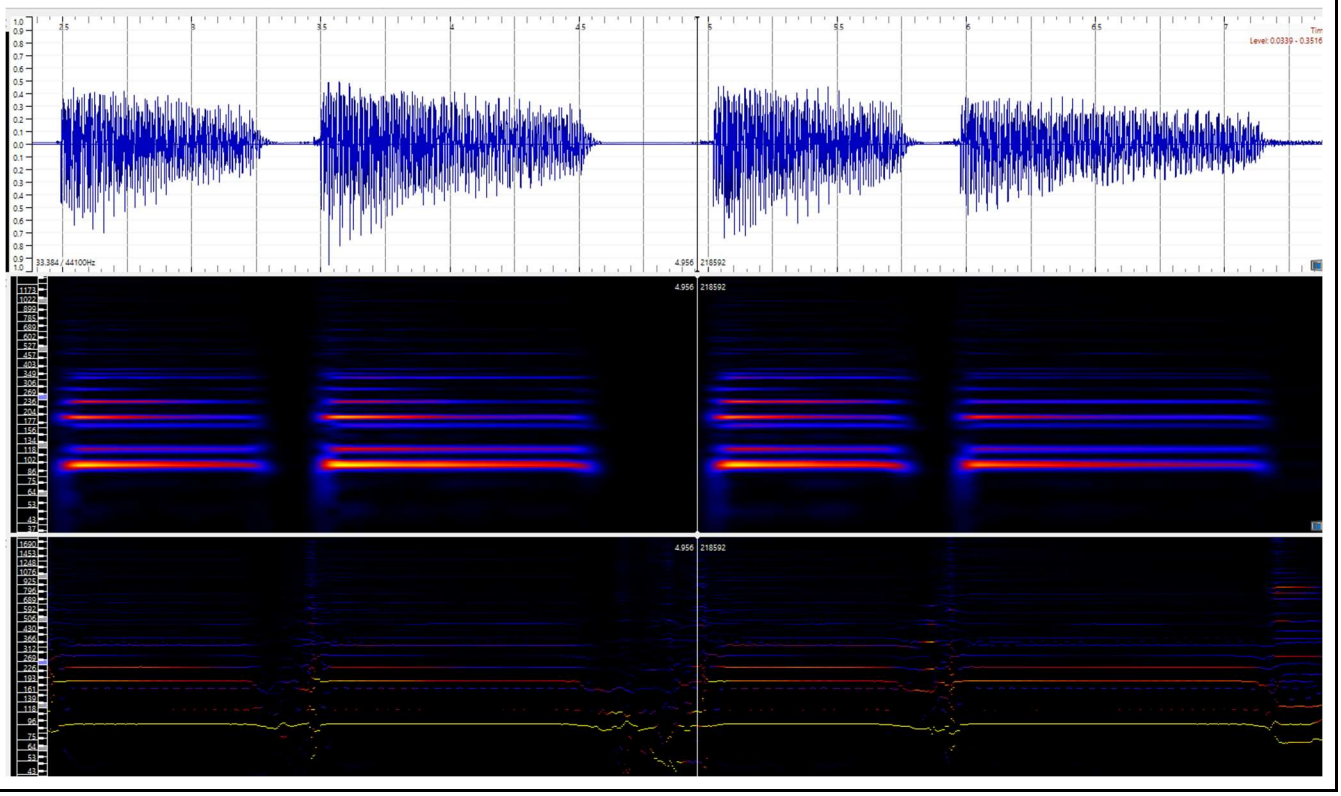
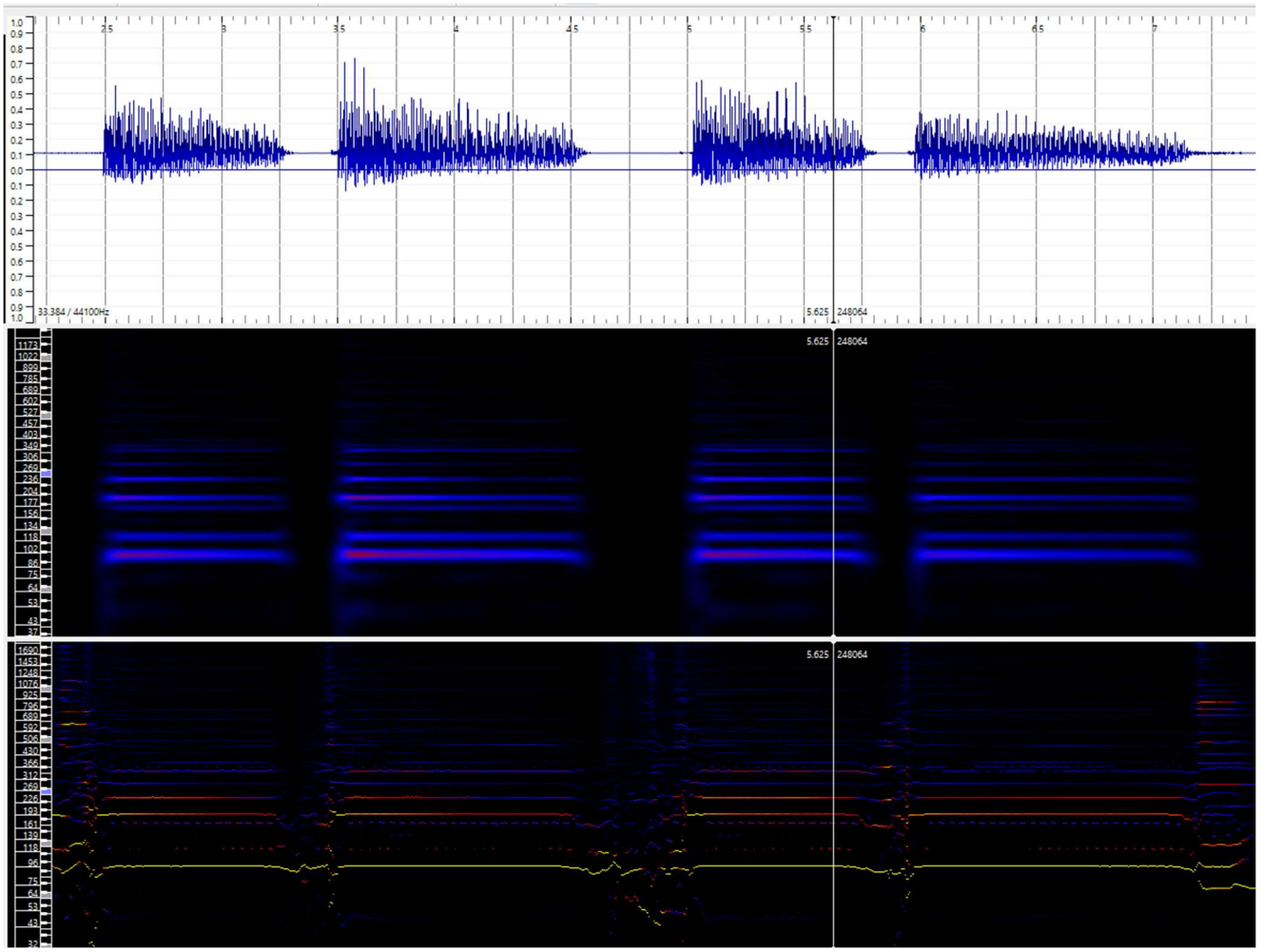
Table 6: Jazz Model Amplifier	 soundtest_jazz_amp.wav

Table 7: Jazz Min Model, 12 layers	 soundtest_jazz_min_model.wav
Table 8: Jazz Mid Model, 32 layers	

	soundtest_jazz_mid_model.wav
Table 9: Jazz Max Model, 64 layers	 soundtest_jazz_max_model.wav

	'Rock" Model: Clean Channel, hard-clipped Overdrive, single-coil pickup

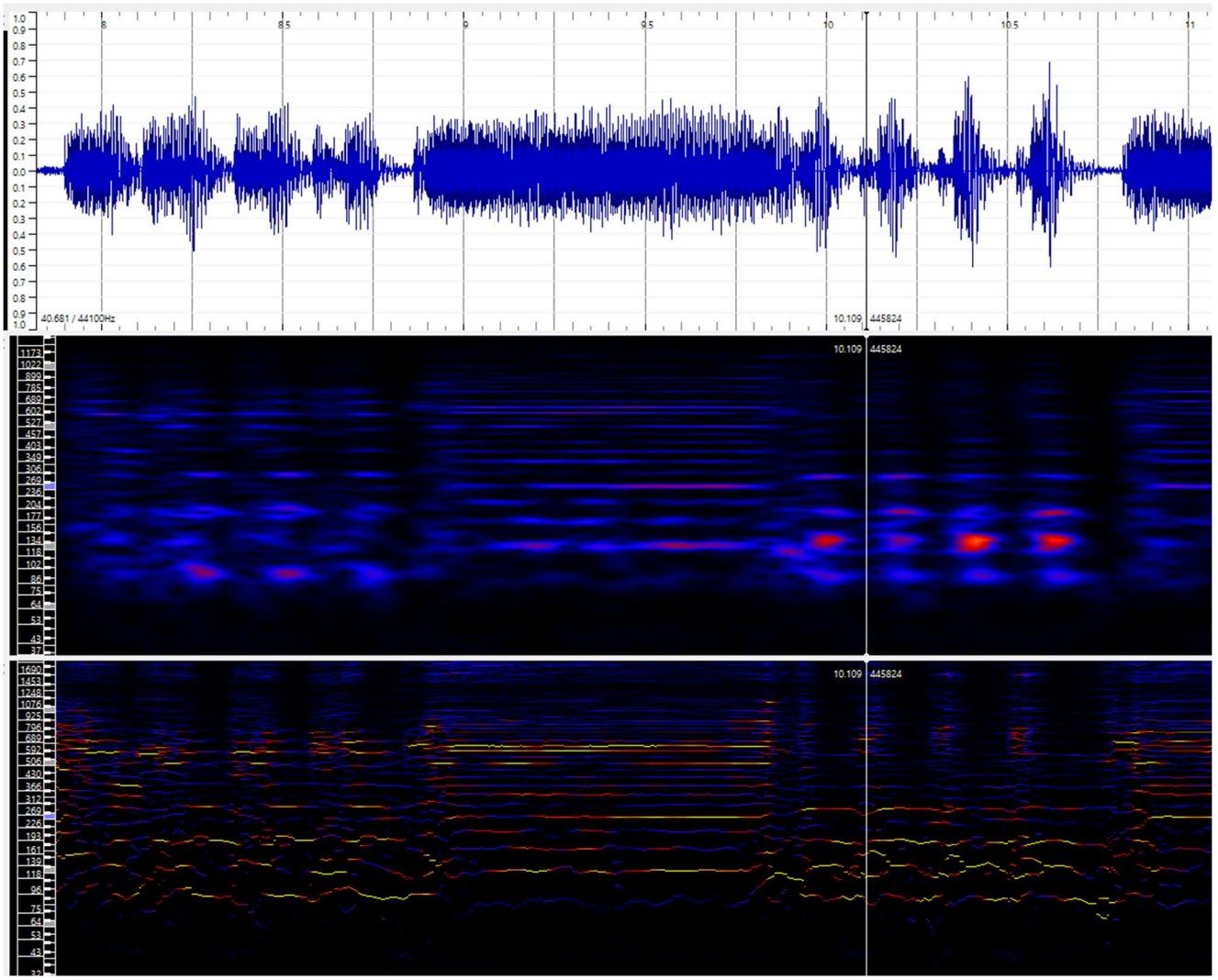
Table 10: Rock Model Amplifier	 soundtest_rock_amp.wav

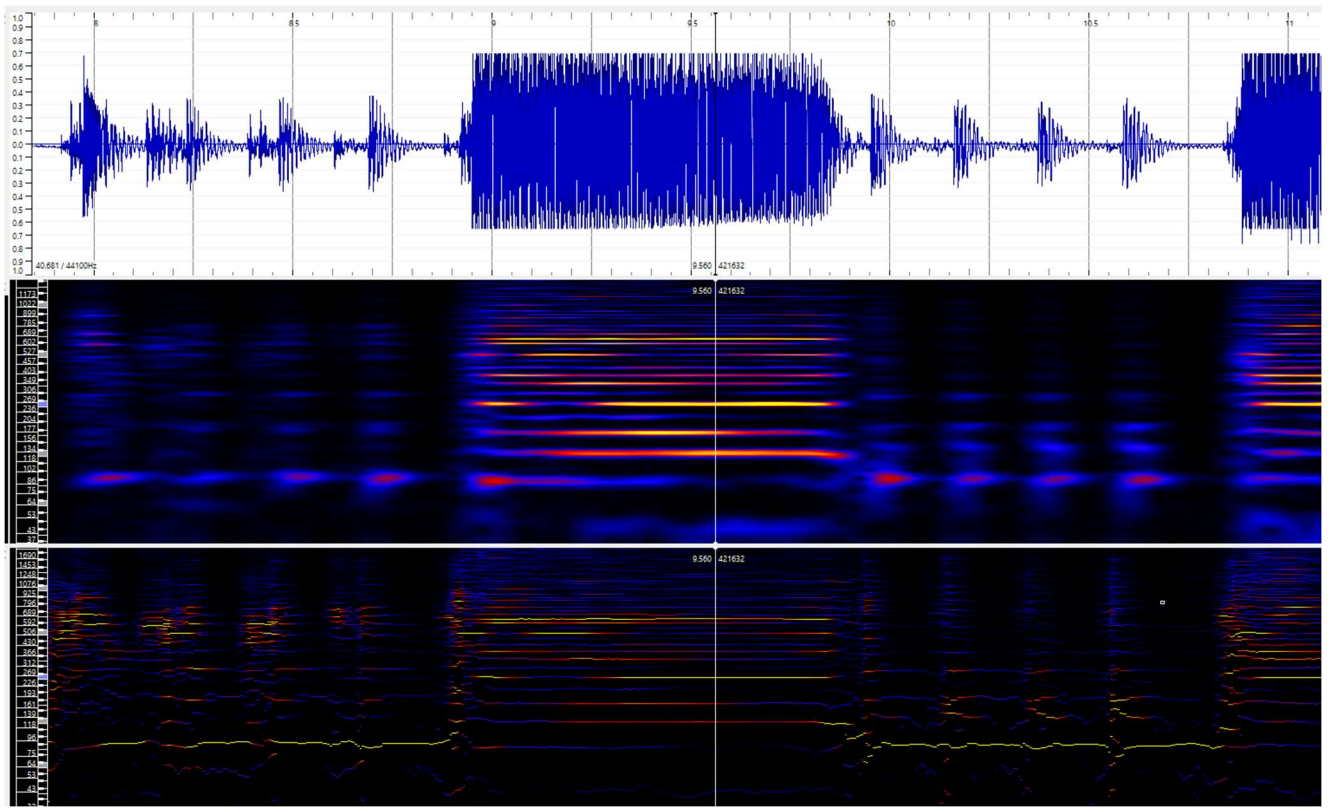
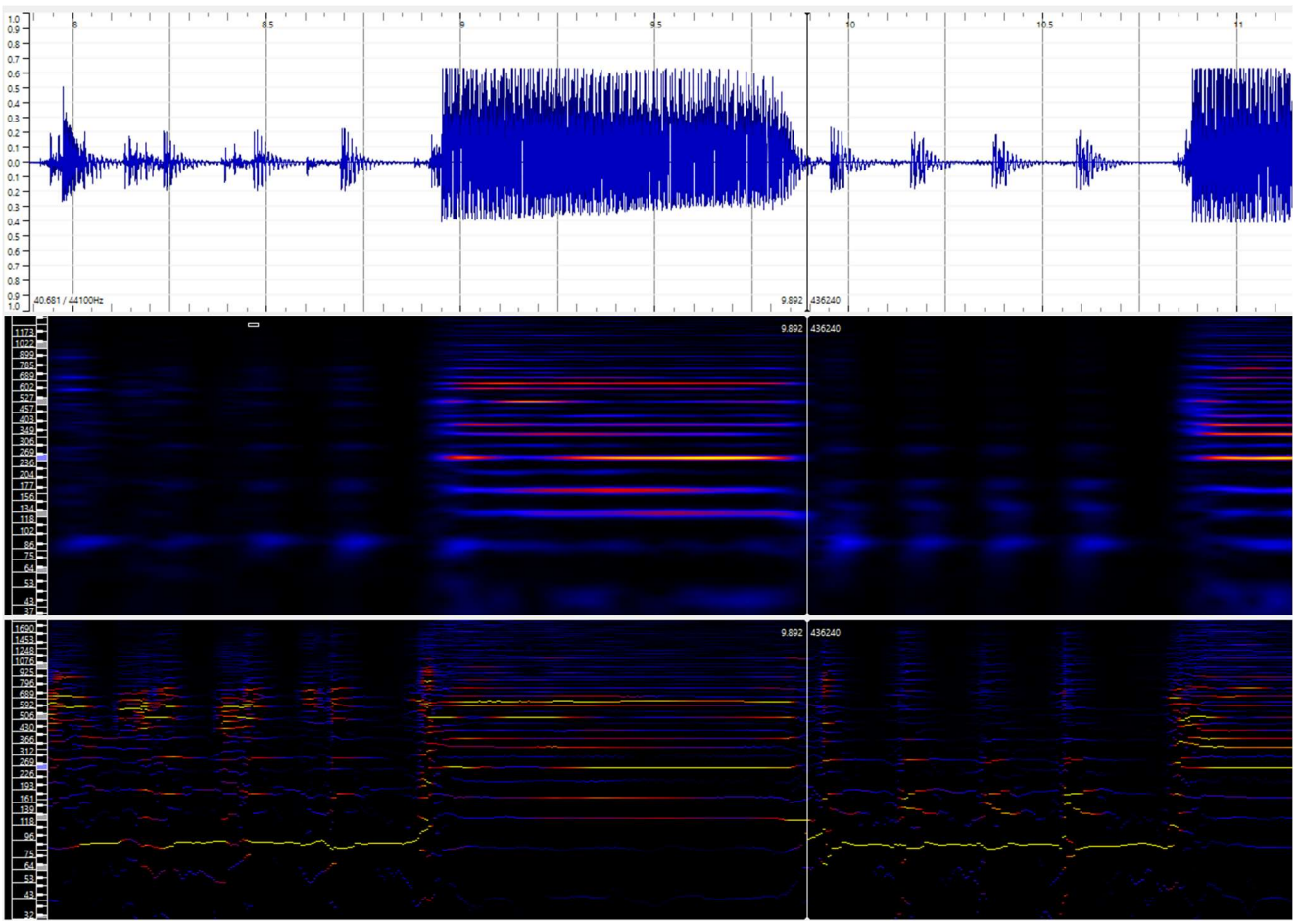
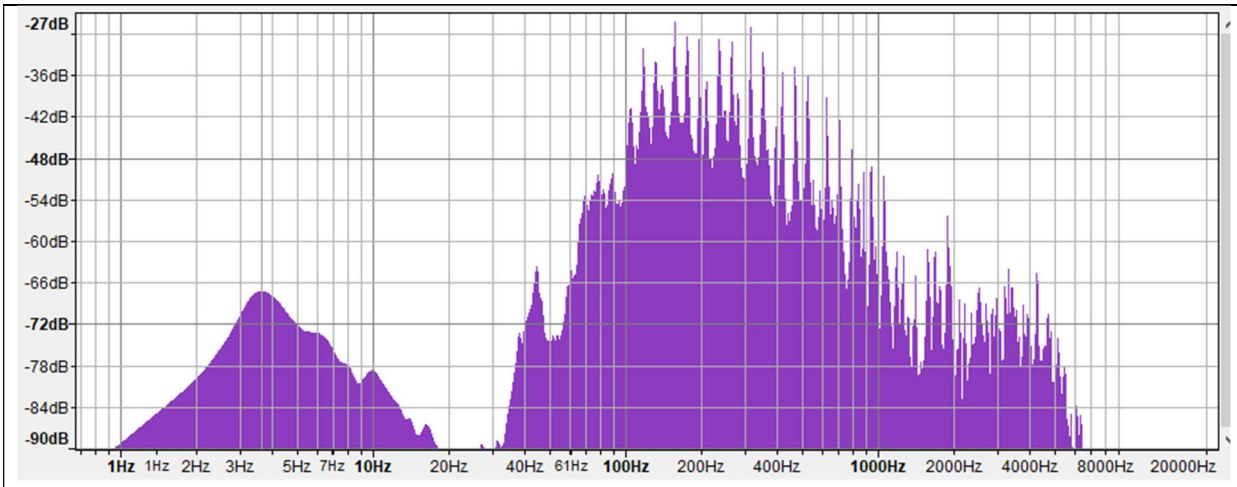
Table 11: Rock Min Model, 12 layers	 soundtest_rock_min_model.wav

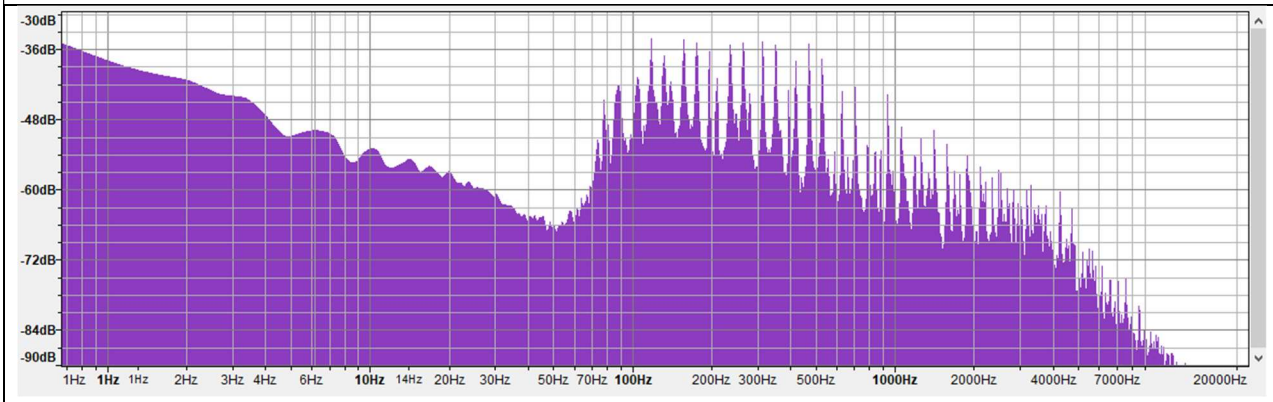
Table 12: Rock Mid Model, 32 layers	 soundtest_rock_mid_model.wav

Long-term frequency analysis was also plotted to look at the overall characteristics of spectra as they appear in the guitar track outputs. Ideally a model without errors would have an identical long-term frequency plot as that of the control source. The models could not correctly learn the low-frequency characteristics of the amplifier. In practice these frequencies which are lower than the fundamental frequency of the lowest playable note on a typical electric guitar are cut out of any guitar track in a digital recording. This pattern, however, is indicative of a problematic long-term memory in the model of lower frequencies. Analysis of these plots shows that more errors happen as we go higher in harmonic frequencies for all models. Errors also dramatically increase in the presence of more powerful harmonics in the higher range

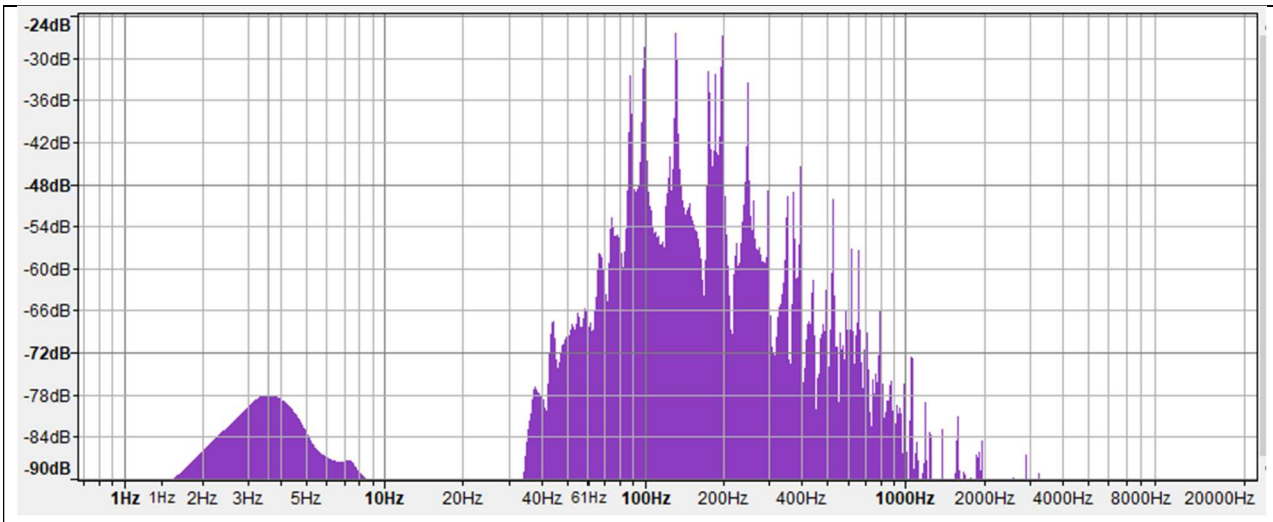
frequencies. The long-term frequency plots show that the amplifier sound typically does not achieve significant power levels of frequencies above 9 kHz. Also the correlation of dynamics are not consistent across different models hyperparameter settings.



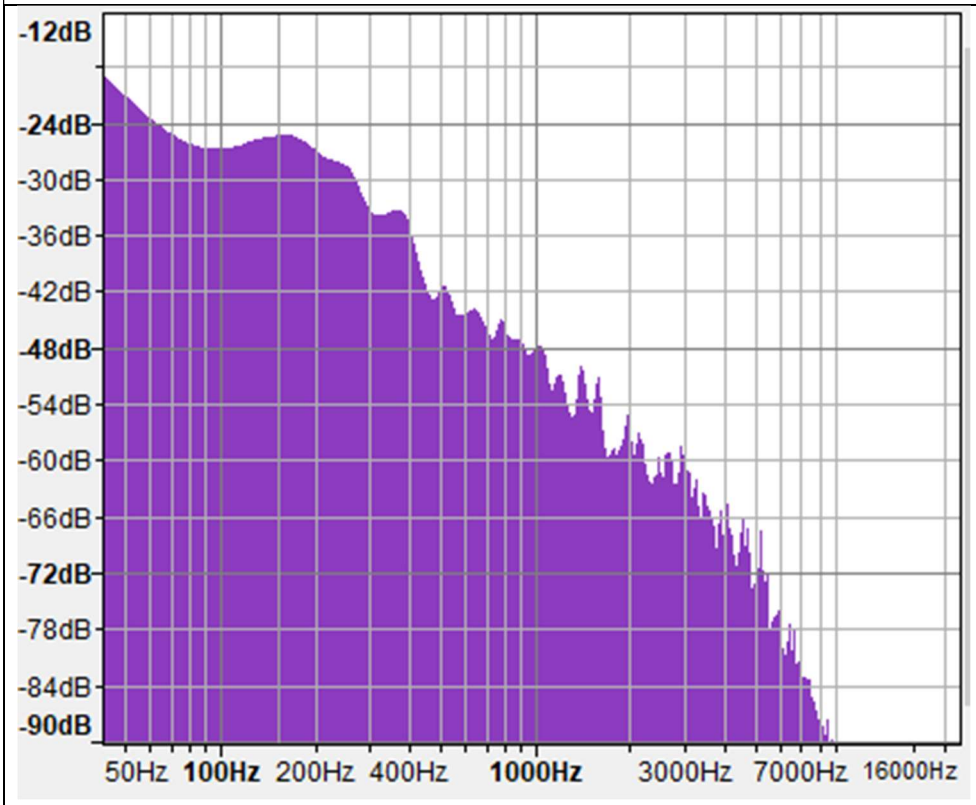
Blues Amplifier



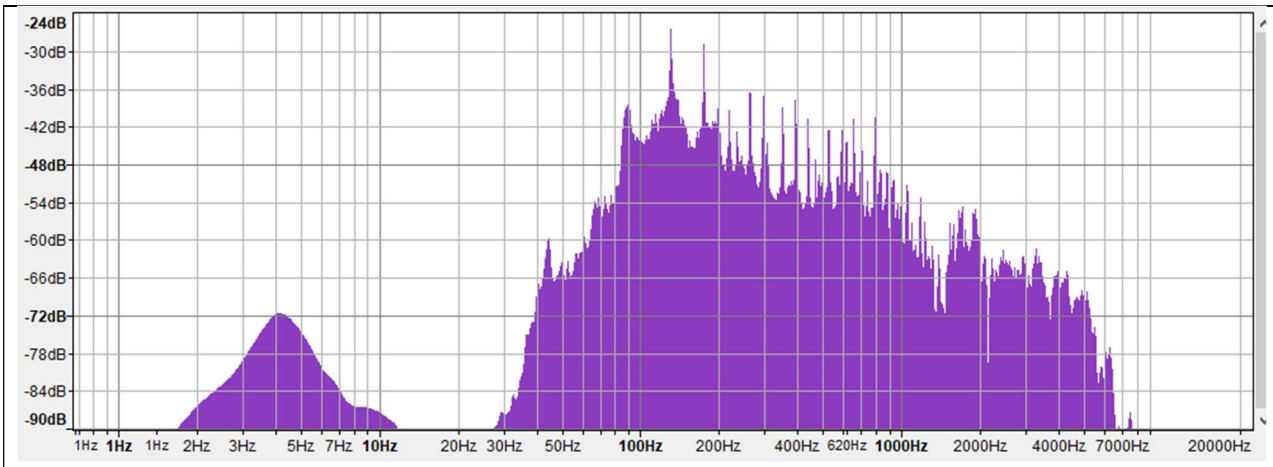
Blues Minimum Model



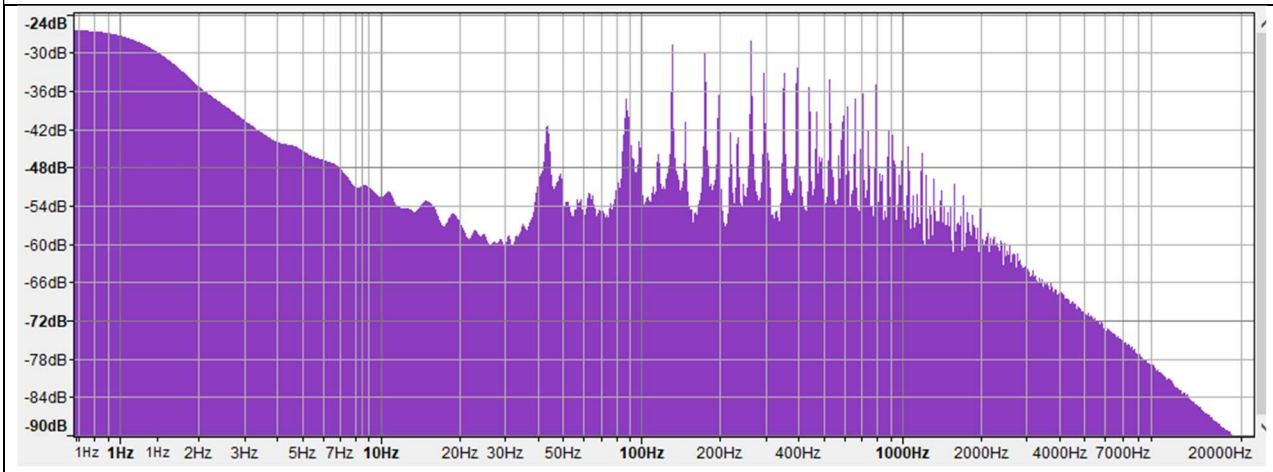
Jazz Amplifier



Jazz Maximum Model



Rock Amplifier



Rock Mid Model

Figure 7: Long-term spectral analysis plots of guitar amps and a corresponding model. Correspondingly more errors happen at increasing harmonic frequencies or with more powerful harmonics at higher frequencies. The model is also unable to learn low frequency patterns. The model does learn a mid-range frequency harmonic pattern somewhat similar to the characteristics of the amplifier.

8.2 Audio engineering based observation

Signals were also analyzed based on the quality of testing output on each model through the use of audio monitoring. For audio monitoring, a pair of high quality speakers were placed within the hearing field of each ear, 5 feet away. Musicians and audio engineers are trained to differentiate subtle differences in timbre from the sound of audio coming from monitors. Hundreds of sample models were trained with different settings and hyperparameters to find a suitable set of models for listening tests. The process of the audio engineering based listening analysis led to the conclusion of poor quality from input data that was too varied, as well as the need to phase correct. Audio evaluation also confirmed that high levels of distortion could not be modeled with satisfying results for any hyperparameter training or other optimization.

8.3 Subjective Listener Tests

On October 15th, 2021 I presented to a group of music students at University of Missouri by playing music samples with guitar sounds from both the amplifiers and models. They sampled the mixed sessions as well as isolated guitar tracks. The music students were chosen as professional listeners who are familiar with listening to subtleties in a musical timbre. The students were offered to take an online ranking test using the WebMUSHRA ranking system. [7] Three students completed the online ranking, and the results are shown. Generally, the models for the Rock track were rated very poor. However, some models, particularly from the Jazz session, suited the musical preferences of some listeners as better sounding than the original amplifier based sound, especially within a mix. The scores representing a general ‘quality’ criteria were highly dependent on the particular tastes of the listener.

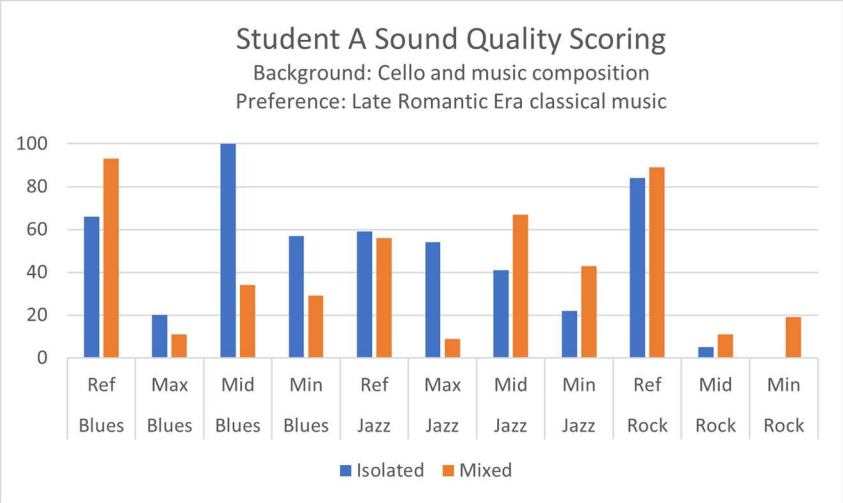


Table 13: Online Sound Quality Scoring of Experimental Output for 'Student A'

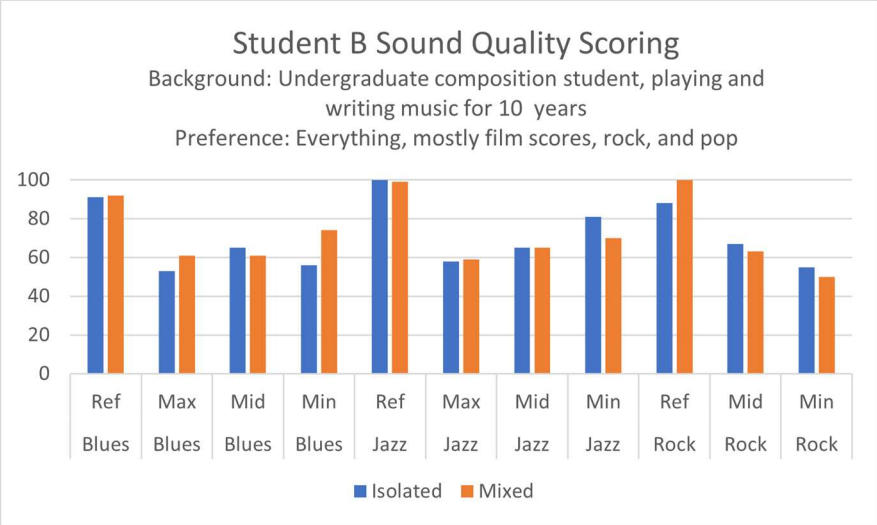


Table 14: Online Sound Quality Scoring of Experimental Output for 'Student B"

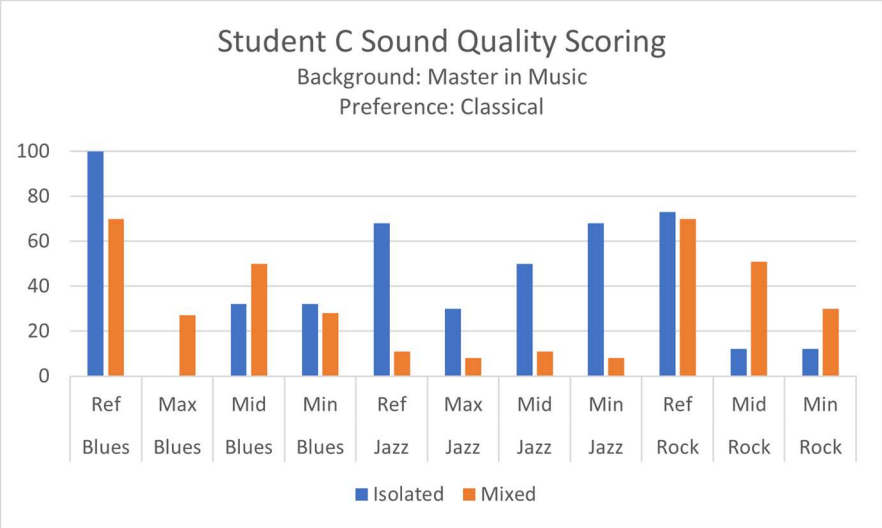


Table 15: Online Sound Quality Scoring of Experimental Output for 'Student C'

CHAPTER 9: CONCLUSIONS

9.1 Differences from previous experiments

This experiment introduced changes to the previous experiments discussed in literature whose effects on performance suggests the WaveNet model has limited applications for replacing real amplifiers. The previous experiment testing PedalNet was in fact very limited in its scope. The previous literature experiment used input data that was very rigid in its expression of dynamics and playing technique. It also did not use the amount of distortion characteristic in modern guitar playing tones. The models in the previous literature were tested on components of an overall signal chain whose distortion characteristics are much simpler than a complete signal chain tested in this experiment.

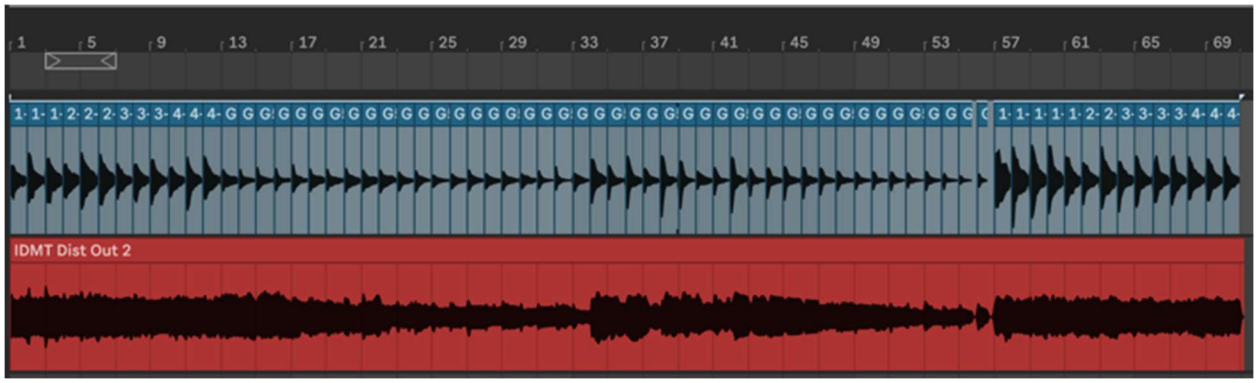
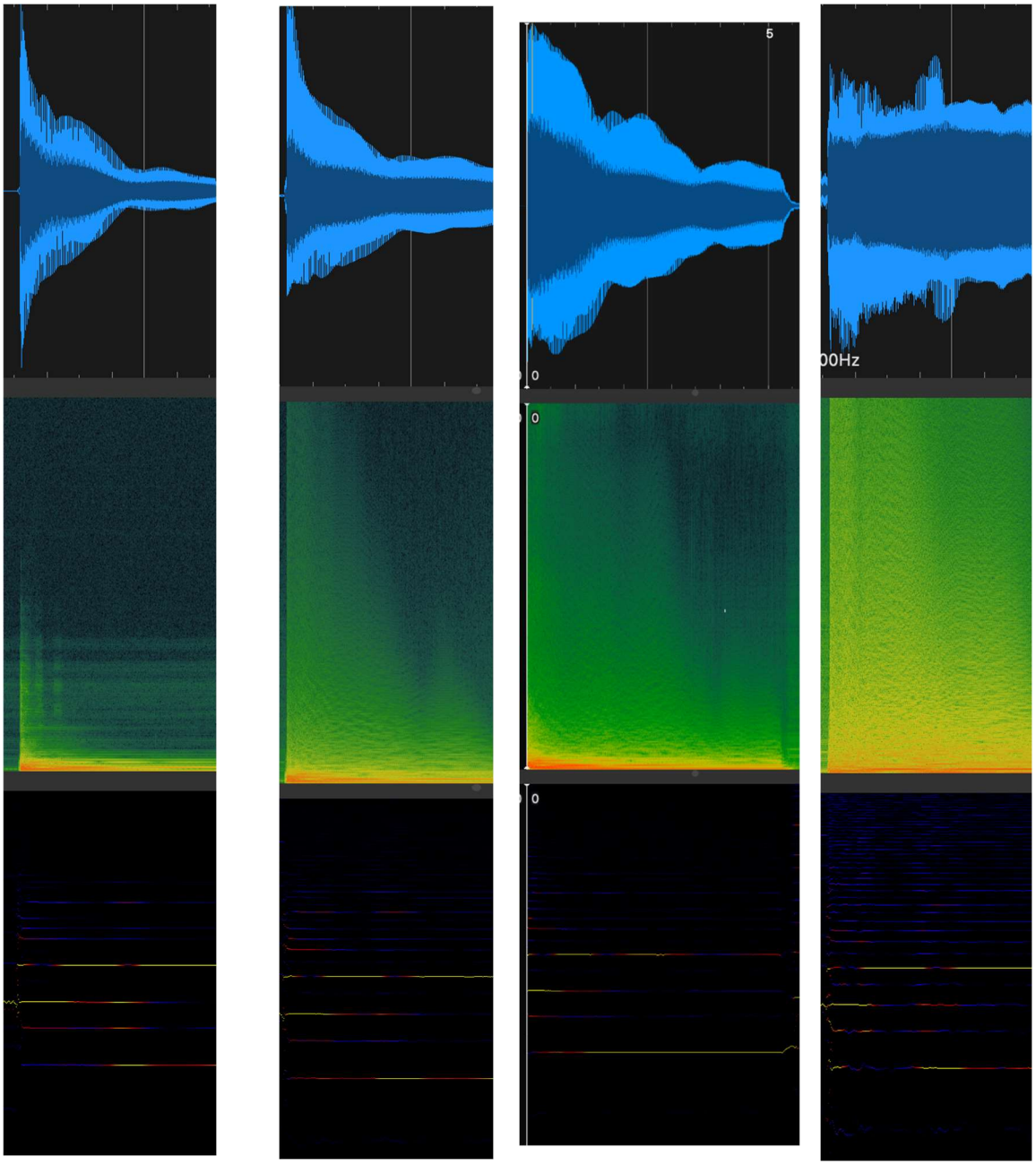


Figure 8: Loaded sound samples used in previous experiments alongside sample amplifier output shows general uniformity of previous experiment training data



Dry signal	Maxon St9Pro+ Tubescreamer Overdrive pedal	Ibanez Ts9 Tubescreamer Overdrive model	Boss DS1 Distortion pedal

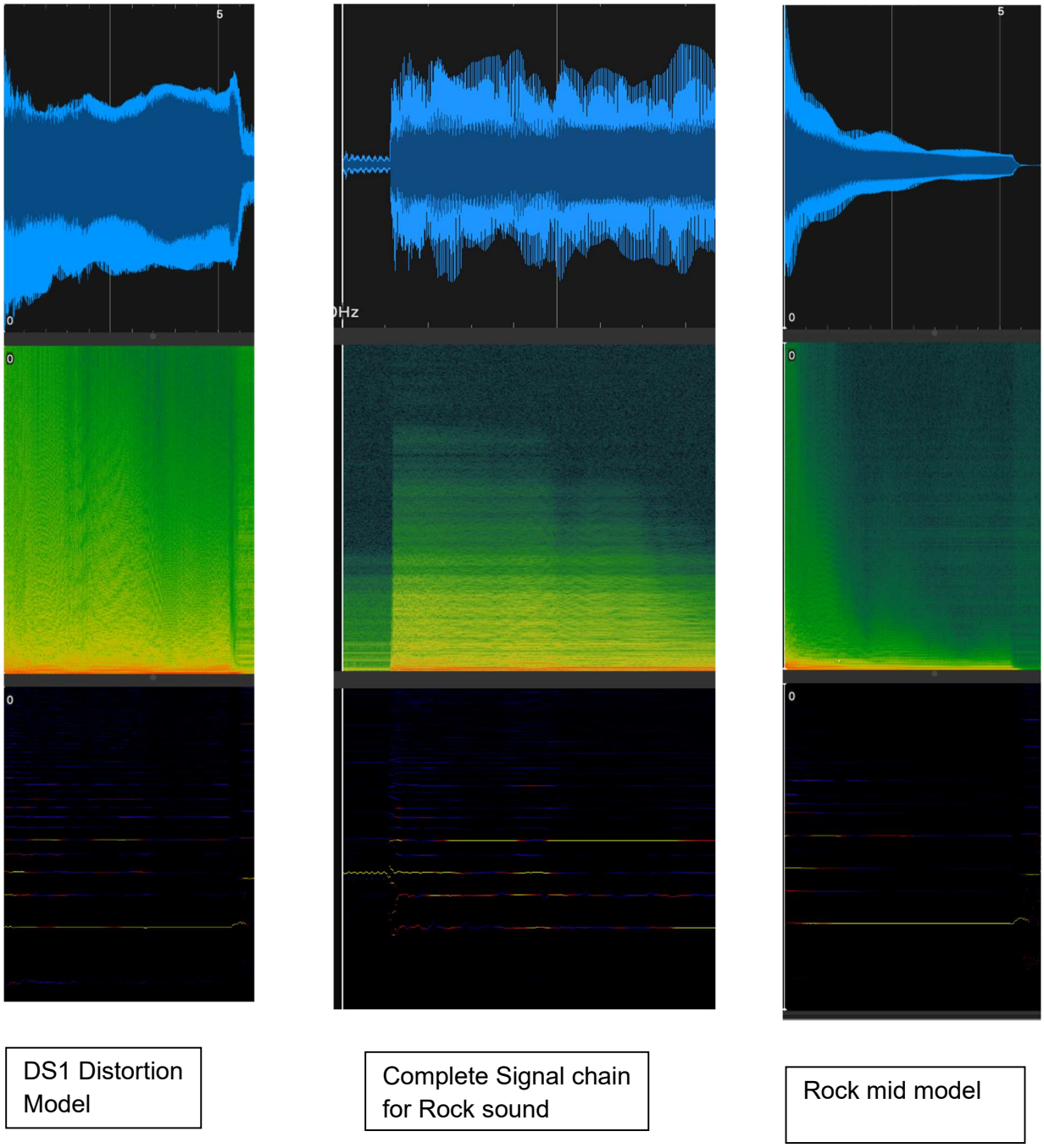


Figure 9: Seven figures show a single power chord played on a guitar. The first is the dry signal. The second four show the effects of solid-state effects 'pedals' designed as signal treatments

before the signal enters the amplifier. These effects were modeled with success in previous literature. The models are freely available online. [9] The last two figures show outputs of a complete electric guitar amplifier signal chain with full harmonic distortion and a corresponding model created for this experiment. No WaveNet type model was successful replicating the dense harmonics even at the highest hyperparameter settings of a complete signal chain.

9.2 Sound qualities captured by model

1. Dynamics

The model learned the relationship between the dynamics of the input and the output training data very well. This meant that the training data had to be normalized in order to facilitate an equivalent relationship between the input and output dynamics. If the model was trained on data where the dynamics between the input and output were distinctly different in average amplitude, the model would follow suit in making a signal louder or quieter.

2. Guitar playing technique

The model was found to have been able to replicate almost any guitar playing technique even without the technique specifically available in the training data. This includes pinch harmonics, palm muting, and pitch bends.

3. Added warmth, roundness of amplifier gain

The model was successful in adding the designed and desired effects of a signal chain through the amplifier circuit that did not involve distortion. Vacuum tubes are known to add a warmth and round-ness to a dry sound and the effects were very apparent even in the minimum models.

9.3 Major limitations of model

1. Too much input variation leads to underfitting

As mentioned earlier, too much input variation leads to underfitting the output of the model. This was not reported in previous literature but discovered through the experiment. This is important because the design of electric guitars may vary substantially to include different woods, construction types, tunings, and electronics.

The most important element in terms of input variation is the output power of the pick-up coils when taking input and going to the amplifier. This is because the output power of the pick-ups can be detected by plugging a guitar directly into a recording interface. However the amplifier will amplify this signal which is then recorded and compressed into a more uniform output. It is difficult for the model to learn the fundamental harmonics of an amplifier to the input when the dynamics are so varied in input.

2. Inaccurate rendering of pronounced distortion

The amplifier and input settings with the highest amount of distortion were uniformly judged as poor quality. The distortion of the models came out sounding like static in the high frequencies. Transient attacks were too short. The transients in the middle frequencies of a distorted sound also did not retain the specific character of the amplifier model for which it is marketed as possessing. This difficulty in learning high amounts of distortion was not reported in the previous literature. However this experiment utilized a signal chain which introduced a much greater presence and complexity of distortion than any signal transformation previously modeled. In particular, the amplifier used was designed specifically for high-gain distortion. This experiment also increased output in the power section of the amplifier, another distortion inducing amplifier feature not previously discussed. Also the recording chain used a studio compressor unit known to introduce desirable harmonics in its compression operation. The recording chain of output signals used in previous literature did not use the complete studio recording setup.

3. Increasing inaccuracy in correspondingly higher harmonics

Similar to the problematic rendering of more pronounced distortion harmonics, the long-term spectral plots in the model outputs show that there are more errors in higher frequencies

than in the fundamental frequencies of the guitar signal. This may be caused by higher frequencies generally having less time steps with which to be predicted.

9.4 Suggested Model

Based on experimental outcome, a model is suggested for future experimentation. These suggestions mostly involve a more complex model with a greater number of parameters for capturing non-linear relationships in distortion. This would most likely require more training data than the approximately 3-4 minutes of training data used for all of the PedalNet models.

1. Finer sample rate to capture higher harmonics

The first suggestion for a future model is to train a model on an audio signal which has more data per unit of time. My suggestion would be to train a model at 96 kHz which is a rate that is available on affordable home studio digital interface equipment. The main purpose of this is to allow the model more ability to learn more detailed relationships between many specific harmonic frequencies on a model.

This hypothesis is counterindicated by the low energy levels of high frequencies above 7k put out by the guitar amplifier in my model. If there is not as much high frequency energy anyway then the existing sampling rate may capture all relevant frequencies with adequate resolution. However I believe because of the poor results in the more distorted sound settings that a higher sampling rate for training data would allow for more accurate modeling of non-linear relationships between the harmonics of the amplifier that carry musical value to the tone.

2. Fully or Partially Connected Layers with no dilations values to capture transients and complex non-linearities

The second and most important suggestion for a future model specific to electric guitar amplification is to have three or more fully or partially connected layers with no dilations

between time steps looking back in the neural network model. I believe this will improve performance for two reasons. First, because the main part of the tone we are interested in is the transient signals happening rapidly at the beginning of the sound. Second, because the elements of a guitar string being plucked can be decomposed into four very short intervals which the model could learn. Both of these observations are meant to argue that a WaveNet model is not exactly suitable for the case of an electric guitar. WaveNet was designed with a human voice in mind. Human voices articulate speech phonetic units much slower than an electric guitar generates a musical tone.

The elements of a guitar tone can be divided into four parts according to physical modeling of instruments. The four parts are attack, decay, sustain, and release. The attack, decay, and release patterns of an electric guitar happen very rapidly. Sustain is a portion of the sound envelope that is modeled as repetitive and unchanging. So a long sustain could arguably be decomposed into a recurring pattern of the same tone happening over time. Since each of these parts of the sound envelope are essentially different elements and distinguishable, a sophisticated enough machine learning model could emulate each of the envelope parts without having to interleave between another. This means the receptive field for a successful model can be shorter than that provided by WaveNet.

Based on my hypothesis that a shorter receptive field is needed to capture the essential elements of a guitar timbre versus a voice, it might make sense to attempt a model involving fully connected layers or a regular convolutional model instead of the dilated convolutional model of WaveNet. An alternative to WaveNet would similarly be structured as a many-to-many recurrent neural network. Instead of the dilated convolutional structure of WaveNet to capture long-term dependencies, the alternative solution would have two or more fully connected or

convolutional layers looking back at previous time steps to capture long-term dependencies. Fully connected layers are more computationally intensive to operate, but the connectedness between layers could more easily capture subtle harmonics that discerning musical tastes might notice. Specialized hardware could also be designed to execute the model in real time.

Some recent work suggests fully connected layers are actually counterproductive to more sparse architectures for certain tasks. [8] The benefits of sparseness can be accomplished with regular convolutions or other type of less sparse connectivity than full connectedness between nodes.

This hypothesis takes into consideration that the models tested in this experiment failed to create the relationship in low frequencies below the fundamental frequency of the guitar for all amp settings. Multiple complete layers in a recurrent model would be able to retain more information on low frequencies than the dilated model.

3. Time forward model dependencies

Because of the nature of electron movement in overdriven electronic components used to generate musical distortion, time forward dependencies in a model would help capture some of the characteristics of this distortion. In op-amps, transistors, and vacuum tubes the distortion characteristics are created when the component is overloaded with electrons to the point that the flow of electrons slows or in some cases starts pushing back against the current going through the component. My hypothesis is that this behavior could be captured with time forward model dependencies. For a model, this would mean that at time step t+n of the input we render some time step t where n is the number of steps forward we wish to have. This feature could possibly be combined with fully connected layers to capture nuanced musical nonlinearities that may be overlooked by models more concerned with mass performance or simplicity.

4. Introduce loudness metrics into model parameters at each time step

Another hypothesis for increasing model quality is to condition each time step on a parameter or multiple parameters measuring loudness in a particular window. Loudness, K-weighted, relative to full scale (LKFS) is a measurement which takes into consideration the psychoacoustic measure of perceived loudness. This measure uses what is called an equal-loudness contour which represents the perceived loudness of varying frequencies at equivalent decibels. This should allow the model to train against a specific psychoacoustic property.

CHAPTER 11: REFERENCES

1. Deep Learning, Ian Goodfellow, Yoshua Bengio, and Aaron Courville, 2016, Cambridge MA MIT Press

2. Springer Handbook of Systematic Musicology, Bader Edition, Editor Rolf Bader, Springer-Verlag 2018,

    1. Construction of Wooden Musical Instruments, Chris Waltham, Shigery Yoshikawa
    2. Music Studio Technology, Robert Mores
    3. Perception of Timbre and Sound Color, Albrecht Schneider

3. Inside Tube Amps, Dan Torres, 2016

4. The Mixing Engineer's Handbook, Fourth Edition, Bobby Owsinski, Bobby Owsinski 2017

5. WaveNet, A Generative Model for Raw Audio, https://arxiv.org/abs/1609.03499, Aaron VanDenOord et al.2016

6. Real Time Guitar Amplifier Emulation With Deep Learning, Alec Wright et al., 2020, Appl. Sci. 2020, 10(3), 766; https://doi.org/10.3390/app10030766

7. Towards the Next Generation of Web-based Experiments: A Case Study Assessing Basic Audio Quality Following the ITU-R Recommendation BS.1534 (MUSHRA), Michael Shoeffler et al., 2017, https://www.audiolabs-erlangen.de/resources/webMUSHRA

8. Rethinking Full Connectivity in Recurrent Neural Networks, Matthis Van Keirsbilck et al., 2019, https://arxiv.org/pdf/1905.12340.pdf

9. WaveNetVA Code Repository, Eero-Pekka Damskägg,

https://github.com/damskaggep/WaveNetVA