PERCEPTUAL EVALUATION AND GENRE-SPECIFIC TRAINING OF DEEP NEURAL NETWORK MODELS OF A HIGH-GAIN GUITAR AMPLIFIER

Will J. Cassidy and Enzo De Sena,

Institute of Sound Recording
University of Surrey
Guildford, United Kingdom
willcassidy7@yahoo.com | e.desena@surrey.ac.uk

ABSTRACT

Modelling of analogue devices via deep neural networks (DNNs) has gained popularity recently, but their performance is usually measured using accuracy measures alone. This paper aims to assess the performance of DNN models of a high-gain vacuum-tube guitar amplifier using additional subjective measures, including preference and realism. Furthermore, the paper explores how the performance changes when genre-specific training data is used. In five listening tests, subjects rated models of a popular high-gain guitar amplifier, the Peavey 6505, in terms of preference, realism and perceptual accuracy. Two DNN models were used: a long short-term memory recurrent neural network (LSTM-RNN) and a WaveNet-based convolutional neural network (CNN). The LSTM-RNN model was shown to be more accurate when trained with genre-specific data, to the extent that it could not be distinguished from the real amplifier in ABX tests. Despite minor perceptual inaccuracies, subjects found all models to be as realistic as the target in MUSHRA-like experiments, and there was no evidence to suggest that the real amplifier was preferred to any of the models in a mix. Finally, it was observed that a low-gain excerpt was more difficult to emulate, and was therefore useful to reveal differences between the models.

1. INTRODUCTION

Analogue vacuum-tube guitar amplifiers are still valued in the audio community, despite being heavy, expensive, and high-maintenance. Historically, several methods have been proposed to emulate vacuum-tube amplifiers [1], including white-box models such as transient modified nodal analysis and wave digital filters [2], and block-oriented grey-box methods such as the Wiener-Hammerstein topology [3]. More recently, advances in deep neural networks (DNNs) have seen promising results compared to traditional approaches [4, 5, 6]. Neural networks are particularly well suited to the black-box modelling of the complex and non-linear internal operations of a guitar amplifier, where training can be performed based on input data (a direct-injected guitar signal) and output data (the distorted, amplified signal) alone. In the literature, DNN models are normally evaluated using objective accuracy measures such as error-to-signal ratio [7, 4] and mean square error based metrics [8, 9, 10]. Subjective accuracy measures have also been seen in the works of [4] and [5], where subjects were asked to rate models in terms of how accurately they approximated the timbre of the reference, and in terms of perceived similarity to the reference, respectively. However, the 'realism' and 'preference' of DNN amplifier models has not been studied to the same extent - is the accuracy of a guitar amplifier model as important as its realism? Also, could a model be preferred to the real amplifier? This paper aims to assess the subjective performance of two popular DNN topologies, namely a long short-term memory recurrent neural network (LSTM-RNN) and a WaveNet-based convolutional neural network (CNN), modelling a popular high-gain vacuum-tube guitar amplifier, the Peavey 6505.

Guitar amplifiers are often specific to certain genres of music. This is especially so for high-gain amplifiers, which are commonly used in heavy rock and metal. Furthermore, certain types of guitar hardware are used more than others, as well as certain playing techniques. Several guitar recordings datasets are publicly available, including the Fraunhofer Institute for Digital Media Technology (IDMT) guitar and bass datasets [11, 12]. These recordings include a range of general techniques, notes and guitar types, and have been used in [10], [8] and [13] to train DNN amplifier models. As Parker *et al.* point out [14], the state-space of an audio system may require certain inputs in order for the target characteristics to be learned effectively, such as for nonlinearities that only occur above a magnitude threshold. On this basis, it is hypothesised in this paper that DNN models of high-gain amplifiers should be trained using data tailored to the target device. In this paper, the IDMT dataset is compared to two genre-specific training files, focused on rock and metal styles, respectively.

The paper is organised as follows. Section 2 describes the target system, i.e. the amplifier and loudspeaker cabinet chain. The DNN models used in this work are then detailed in Section 3, the training of which is outlined in Section 4. The methodology and results of the listening experiments are presented in Section 5, and the results are discussed in Section 6. The main conclusions are summarised in Section 7 with suggestions for further work.

2. TARGET SYSTEM

The target system consists of a guitar amplifier and a loudspeaker cabinet. Only the guitar amplifier was modelled as part of the DNNs, while the loudspeaker cabinet was modelled separately [13], as a linear time-invariant (LTI) system.

The selected target amplifier is the high-gain vacuum tube Peavey 6505, a popular choice in metal recording. High-gain amplifiers usually consist of a preamplifier stage with around 3-7 small vacuum-tubes, commonly 12AX7 dual-triodes, which provide the majority of the non-linear signal distortion [2, 15]. For high-gain amplifiers, a 'drive' parameter applies gain to the input signal before this stage to drive the preamplifier tubes, thus in-

creasing the total harmonic distortion (THD) of the system. The preamplified signal passes through a linear tone stack circuit before power amplifier vacuum-tubes provide signal gain to sufficiently drive the loudspeaker cabinet [2]. These power tubes contribute to the linear tonal characteristics of the amplifier, referred to as the 'British' or 'American' tone depending on their model [15].

The selected target loudspeaker cabinet was a Marshall 1960-AV, consisting of four 12-inch Celestion Vintage 30 loudspeakers which were also used in [16]. The impulse response (IR) of the loudspeaker cabinet was measured using a 30-second long exponential sine sweep (ESS). The sweep was generated at -6 dBFS which was routed to the line output of a Universal Audio Apollo Twin X audio interface. A QSC RMX850 power amplifier was used to apply clean gain to the ESS signal, the output of which was connected to the matched-impedance input of the loudspeaker cabinet. The loudspeaker response was recorded using a Royer R-121, a professional-grade figure-8 ribbon microphone with a 30–15,000 Hz $\pm$3dB response and very high overload characteristics (135 dB SPL). The microphone was positioned at approximately 20mm off-centre from the dust cap. The output SPL of the loudspeaker cabinet was set high enough to provide sufficient SNR, yet not to the extent where significant cone breakup was introduced.

3. DNN MODELS

Two DNN topologies that have been previously used for guitar amplifier modelling are a feedforward variant of WaveNet and an LSTM-based RNN, both of which are compared in [10], [13], and [17]. The implementation used in this paper for the WaveNet-based CNN is the *PedalNetRT* repository, while the one used for the LSTM-RNN is the *Proteus* repository [18]. *PedalNetRT* modifies the original *pedalnet* repository [19], which was a recreation of the WaveNet-based model from the paper by [7]. The modification uses custom causal padding and reorganises conv1d layers to allow trained models to be saved as .json files, which can be loaded using the audio plugin from the *WaveNetVA* repository [20].

The *Proteus* project consists of an audio plugin built using *RTNeural*, a realtime C++ inferencing engine [21]. The plugin can load models trained using the *Automated GuitarAmpModelling* respository [18], forked from Wright's repository [22], which is an implementation of the LSTM-RNN network used by Wright *et al.* [13] in their modelling of the Blackstar HT-1 amplifier and the Big Muff Pi pedal.

Despite both of these models being capable of conditioned training, where the effects of varying a parameter such as drive can be learned, the models in this paper were designed to be a 'snapshot', i.e. a model of the amplifier with fixed parameters, since this was sufficient for the scope of the experimentation.

The LSTM-RNN models used the hyperparameters recommended by [18] for medium to high-gain amplifier emulation. This used an LSTM hidden size of 40 as required by the *Proteus* audio plugin, no pre-emphasis filtering, one recurrent block, and a skip-connection. The WaveNet-based models in this experiment used the default hyperparameters from *PedalNetRT*, i.e. 9 layers, 4 convolution channels, a kernel size of 3 and a batch size of 64. While this is lower than what was suggested by Wright *et al.* [7], these hyperparameters result in a running complexity closer to LSTM-RNN. It is acknowledged that this hyperparameters configuration does not represent the full potential of the WaveNet-based CNN, and therefore the two DNN topologies are not compared directly in this work.

4. TRAINING

This section details the process of producing the direct inject (DI) and amplifier signals for three training sets: a general dataset, an existing genre-specific dataset, and a proposed genre-specific dataset. Both DNNs introduced in Section 3 were trained on each training set, using back-propagation with a loss function based on the error-to-signal ratio (ESR). Google Colab was used to train the LSTM-RNN models, and the WaveNet-based models were trained remotely using the University of Surrey High Performance Cluster, utilising the Python preparation and training files provided in the aforementioned repositories by Bloemer [18].

4.1. Existing Training Datasets

The training file used by Wright *et al.* [13] in their emulation of the Blackstar HT-1 amplifier, accessible from [18] is used here. This training set was constructed using excerpts from the Fraunhofer IDMT databases, forming a 5 minute, 40 second file of half bass and half electric guitar. A range of pickup selections and string gauges were used, and the main playing techniques are described in Table 1. This training is henceforth referred to as the *general* training.

Bloemer [18] recorded a set of genre-specific training samples featuring a wider range of techniques and notes than the *general* file, lasting 3 minutes and 31 seconds. The excerpts in this training are more rock-oriented than those of the IDMT database, and the duration was weighted more towards electric guitar than bass. This training dataset is henceforth referred to as the *rock-specific* training, and serves as a middle ground between the general training and the training made specifically for the Peavey 6505.

4.2. Proposed Metal-specific Training

Rock and metal genres of music share many electric guitar techniques, with some aspects being more exclusive to metal such as pinch-harmonics and low tuning. In this paper, a training file is proposed which was created by recording popular metal guitar excerpts, with a focus on more specific metal techniques highlighted by [23], which were not present in the other datasets. These techniques are outlined in Table 1.

The guitars used for the proposed dataset were the Schecter KM-7 MKIII Artist, the Ibanez RG421 with Bareknuckle Aftermath pickups, and the Dingwall NG-2 5-string. When recording each guitar, the output was connected in series to a true-bypass Peterson tuner pedal, a Radial J48 active DI box and the microphone input of a Universal Audio Apollo Twin X interface. Engaging the -15dB PAD (passive attenuation device) on the DI box was necessary for the active guitars as the input transformer was being overloaded, and so it was engaged for all guitars for consistency. The guitar volume/tone potentiometers were first set to 100% (most transparent), and the preamplifier gain was set such that 10dB of headroom was present when palm-muting heavily. The UAD Diezel Herbert amplifier simulator [24] was used for monitoring purposes.This proposed training dataset lasts a total of 5 minutes and 27 seconds and is henceforth referred to as the *metal-specific* training.

4.3. Training Comparison

Table 1 presents musical aspects of the three training sets, where the genre-specific training files can be seen to have a more ex-

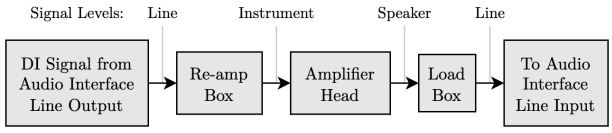

Figure 1: *Signal chain for the amplifier recording process. Signal levels are annotated to highlight the importance of each block.*

Figure 2: *Equipment setup for the recording of the amplifier output.*

tended range of techniques and notes compared to the *general* set. The *rock-specific* training has the widest frequency range when considering the exponential sine sweeps and noise samples at the start of the file.

4.4. Recording the Amplifier Output

The direct output of the Peavey 6505 *LEAD* channel was recorded for each of the three DI training files. The recording chain was set up as per the block diagram in Figure 1, the equipment of which is shown in Figure 2. A Radial X-Amp active re-amp box was used to attenuate the line-level audio interface output to instrument level, with an output impedance of 10kΩ. Typical electric guitar output impedances are in the range of 5-12kΩ [25] - the values of which are expected to be seen by the input of a guitar amplifier. Presenting the correct output impedance is important to ensure the voltage drop across the amplifier is within nominal levels, in order for the amplifier to behave as expected.

The Rivera RockCrusher load box was used to attenuate the high-power output of the amplifier to line level. This is required in replacement of a loudspeaker cabinet, since powering a vacuum-tube amplifier without sufficient load can be damaging [26]. As [13] points out, the type of load may influence the behaviour of the amplifier differently to a loudspeaker cabinet. Therefore, care was taken to select a high-quality reactive load box to act as transparently as possible. The load box output impedance of 560Ω [26] allows for optimal voltage transfer to the line-level input of the Apollo Twin X (rated at 10kΩ [27]). After a preliminary recording, the preamplifier gain was increased to compensate for the voltage loss resulting from headroom provided at earlier stages.

5. SUBJECTIVE LISTENING EXPERIMENTS

Each of the three training sets introduced in Section 4 were used to train the LSTM-RNN and WaveNet-based CNN, resulting in 6 models of the Peavey 6505. Listening tests were conducted to investigate the perceptual preference, realism and accuracy of these models, compared to the real amplifier. The test samples are made

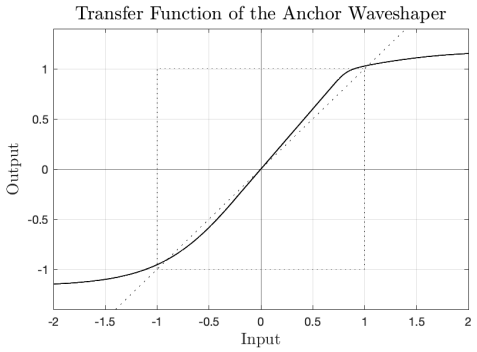

Figure 3: *The transfer function used as a static waveshaper to produce the anchor test samples.*

available on the Institute of Sound Recording's GitHub page[1].

5.1. Test Subjects

A total of 21 participants took part in the first three listening tests. In an anonymous survey, 86% of subjects said they had critically listened to rock or metal music before (i.e. in studio monitoring conditions), 76% had previous experience with vacuum-tube guitar amplifiers, and 81% had experience with hardware or software amplifier simulators.

In the final two listening tests, 16 people took part, all of which had used a vacuum-tube guitar amplifier before, 94% had used an amplifier simulator before and 88% had critically listened to rock or metal music.

All subjects were students of the BSc in Music and Sound Recording course (Tonmeister) at the University of Surrey, all of whom received critical listening training as part of their curriculum.

5.2. Test Excerpts

A range of pickups were used to record 8 guitar excerpts as detailed in Table 2. For each pickup, two excerpts from existing rock and metal songs were chosen with different pitch registers. Playing techniques were also different between excerpts, which included variations of the guitar volume control. Acting as an attenuation device before the amplifier, the volume control can be used to reduce the guitar output level to the 'edge of breakup' (also known as breakup point [9]). These test samples were recorded as 44.1kHz, 16-bit integer linear PCM waveform files, and were each 5-8 seconds in length.

An anchor was produced using a static waveshaper (as in [7]), created using a piecewise transfer function shown in Figure 3. This transfer function is based on the vacuum-tube-like waveshaper designed by [28], and was intended to be distinguishable from the real amplifier due to its simplicity.

The 8 DI guitar excerpts were sent through each of the 6 models, the real amplifier and the anchor waveshaper, resulting in 64 test samples. The models were captured via the *WaveNetVA* and *Proteus* audio plugins in *REAPER*, and the real amplifier output was recorded as part of the process in Section 4.4. The output signals were then convolved with the loudspeaker cabinet IR, and

[1] `https://github.com/IoSR-Surrey/DNNAmplifierDemos`

Table 1: *Information about each of the three training data sets from inspection. The pitch ranges consider the lowest and highest notes played, excluding harmonic techniques.*

Training Data Set	Set 1: General [13]	Set 2: Rock-specific [18]	Set 3: Metal-specific (proposed)
Excerpt Durations	Approx. 10-30s	Approx. 1.5s	Approx. 5-10s
Pitch Range	E1-A#4 (3.5 octaves)	E1-C6 (4.67 octaves)	C1-C6 (5 octaves)
Techniques	Picked bass	Picked bass	Picked bass
	Fingerstyle bass	High-velocity strumming	Pinch harmonics
	Slap bass	Palm-muting	Tapped harmonics
	Strummed dead-notes	High-pitched monophony	Tremolo picking
	Fingerstyle arpeggios	Double-stops	Double-stops
	Monophonic notes	Low-velocity arpeggios	Strummed chords
	Staccato chords	Full-tone bends	Vibrato
	Picked arpeggios	Rapid monophonic picking	Intermod. distortion
	Background tone	Strummed chords	Full-tone bends
		Palm-muted scales	Hammer-ons/pull-offs
		Scales high and low	Fast picking runs
		Natural harmonics	Fast legato
		Vibrato	Volume roll-off
			Heavy palm-muting

equalisation was applied (-9.7dB notch filter at 3.8kHz, Q = 19) to reduce the rate of fatigue of each test subject.

5.3. Experimental Methodology and Statistical Analysis

The experimental methodology involved MUSHRA-style tests [29], which have been previously used in this context [4, 5, 6]. Since this work investigated subjective measures beyond model accuracy, a reference was not used in tests where this would bias the subject's opinion. An ABX test was also used to evaluate model accuracy, which is recommended for the evaluation of smaller differences [30]. The listening tests were conducted inside an acoustically treated room in the Institute of Sound Recording at the University of Surrey. The stimuli were presented to subjects via a *Max/MSP* patch on a 2019 MacBook Pro, monitored over Audio-Technica ATH-M40X headphones. Before the listening tests, each participant was guided through familiarisation, training and blind grading phases. Subjects were made to familiarise themselves with all the unlabelled stimuli and the GUI before conducting the test. During this process, listeners were encouraged to set the monitoring volume to a comfortable level.

The statistical analysis of all MUSHRA-style tests was based on (non-parametric) Friedman tests [31] and post-hoc Wilcoxon pairwise signed-ranks tests. Considering that the paper aims to assess the effect of training within each model, and how well each model performed against the real amplifier, only the following pairwise tests were run: (a) differences between the 3 LSTM-RNN models, (b) differences between the 3 WaveNet-based CNN models, and (c) differences between all models and the real amplifier, for a total of 12 comparisons. The Bonferroni correction was applied to adjust for multiple comparisons.

5.4. Experiment 1 - Preference

The aim of the first experiment was to gauge which amplifier the subjects preferred, be it real or artificial. The test samples were presented in a mix of drum kit and bass guitar to simulate the listening conditions the consumer would experience when judging the guitar recording of a song. The test prompt was worded as: "Rate your preference of the electric guitar in samples A-G".

5.4.1. Methodology

The test used a MUSHRA-style methodology, but without a labelled reference of the real amplifier, so as to account for the possibility of the real amplifier not being the preferred stimulus. Also, an anchor was omitted to reduce the compression of results since the samples appeared to sound very similar. Subjects were asked to rate 7 test samples (i.e. the 6 models and the real amplifier) side by side for 4 different excerpts. Each excerpt was presented on a different page, and each page was repeated once. The scale ranged from -50 to 50 with -10 to 10 labelled as "Indifferent", -50 labelled as "This sounds worse than the others" and 50 labelled as "I prefer this to the others".

5.4.2. Results

The results of the preference test are presented in Figure 4. The mean scores and 95% confidence intervals for each model and the real amplifier were each within the -10 to 10 category, labelled "Indifferent". A Friedman test revealed that there was a statistically significant difference between some of the models ($\chi^2(6, N{=}168)$ = 14.409, p = 0.025). However, post hoc Wilcoxon signed-ranks tests showed that there was no statistically significant difference between the three LSTM-RNN models (i.e. the three different training sets) or between the three WaveNet-based models. Similarly, there was no statistically significant difference between each of the 6 models and the real amplifier.

5.5. Experiment 2 - Realism

This experiment aimed to investigate what subjects believed sounded like a 'real' amplifier given their previous experience of vacuum-tube guitar amplifiers, without a reference. Two MUSHRA-style listening tests were conducted. The first test asked subjects to rate how 'real' the samples sounded, and the second test asked subjects to compare the samples to their previous experience of what a real

Table 2: *Information about each of the 8 excerpts used to form the listening test samples. 'EOB' refers to 'edge of breakup': the lowest of the gain levels. Under 'Song Based On,' the artist is not included for space reasons (full details are provided in the Github repository 1).*

Excerpt	Pickup Model	Pickup Passivity	Song Based On	Tuning	Pitch Register	THD
A	EMG Humbucker	Active	B.Y.O.B. 0:41-0:46	Drob Db	Low	Med/High
B	EMG Humbucker	Active	Tears Don't Fall 0:00-0:06	Drob Db	Mid	Med
C	Fishman Fluence Humbucker	Active	Death Inside 2:21-2:27	Drop Bb	Low	High
D	Fishman Fluence Humbucker	Active	Catalyst 1:31-1:39	Drob Db	Mid	High
E	Fishman Fluence Split-coil	Active	Cry of Achilles 0:32-0:39	Eb Standard	Mid	High
F	Fishman Fluence Split-coil	Active	My Curse 0:00-0:08	Drop C	Mid/High	EOB
G	Bareknuckle Aftermath Humbucker	Passive	My Curse 1:01-1:09	Drop C	Low	High
H	Bareknuckle Aftermath Humbucker	Passive	Buried Alive 4:14-4:20	Standard	High	High

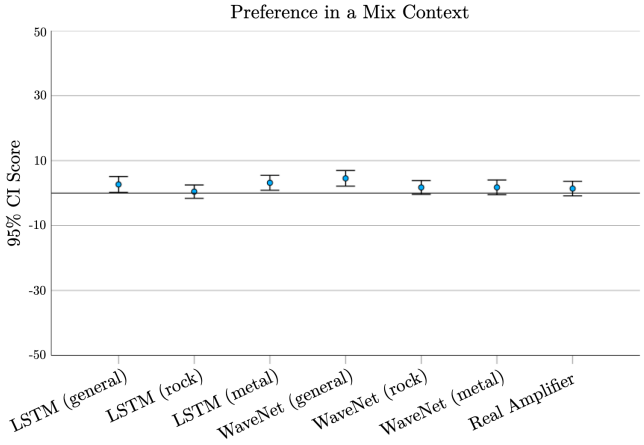

Figure 4: *Means and 95% confidence intervals for the listening test in experiment 1. The results are averaged across all excerpts (C, D, E and G).*

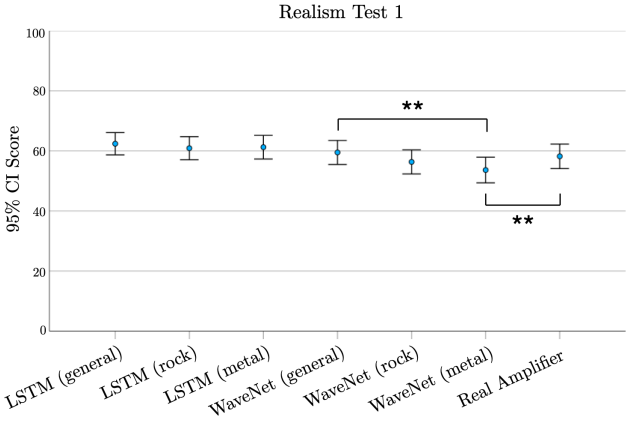

Figure 5: *Means and 95% confidence intervals for the first listening test in experiment 2. Asterisks and bars indicate a significant difference (*: p < .05, **: p < .01 , ***: p < .001 at post-hoc test, Bonferroni corrected).*

vacuum-tube amplifier sounds like. Only subjects that had used a vacuum-tube guitar amplifier before were permitted to participate in the second experiment. A total of 21 subjects participated in the first test, while 16 participated in the second test. These samples were not presented with accompaniment unlike Section 5.4, since realism should not depend on other instruments - using a mix may cause unnecessary masking effects.

5.5.1. Methodology

The first test of this experiment asked subjects to "rate samples regarding how 'real' they sound" on a scale of 0-100 labelled from "This sounds artificial" to "This sounds like a real amplifier". The test consisted of 8 pages, each of which involved comparing the 6 models and the real amplifier using one guitar excerpt as an input. The excerpt was randomly changed for each page, using excerpts A, C, E and G from Table 2 and repeating them once.

The second test asked listeners to "rate each sample based on how similar it sounds to a real vacuum-tube guitar amplifier". On each page of the second test, 8 unlabelled samples were compared (the 6 models, the real amplifier and the anchor). The rating scale was also 0-100, labelled from "Not similar" to "Sounds the same". This was completed for 6 excerpts (A, B, C, F, G and H) and repeated once, resulting in a total of 12 pages.

5.5.2. Results

The results of the first realism test are shown in Figure 5. A Friedman test showed a significant difference between some of the models ($\chi^2(6, N{=}168) = 35.907$, $p < 0.001$), so post hoc Wilcoxon signed-ranks tests were performed. There was no statistically significant difference between the LSTM-RNN models. For the WaveNet-based models, on the other hand, the *general* model was significantly more realistic than the *metal-specific* model (p = 0.0097, adjusted). When comparing each model with the reference, the real amplifier was only significantly more realistic than the *metal-specific* WaveNet-based model (p = 0.0034, adjusted).

Figure 6 shows the mean realism scores of the second test, where a Friedman test also returned statistically significant differences ($\chi^2(6, N{=}192) = 16.412$, $p < 0.012$). Post hoc Wilcoxon signed-ranks tests showed no statistically significant differences between the LSTM-RNN models. Within the WaveNet-based models, the *general* model was significantly more realistic than the *metal-specific* one (p = 0.0269, adjusted), as was seen in the first test. For all of the models there was no statistically difference from the real amplifier.

To examine the effects of the model and excerpt on the mean realism scores of the second test, Friedman tests were run for each of the 6 excerpts used. For the results of excerpt F, a Friedman test

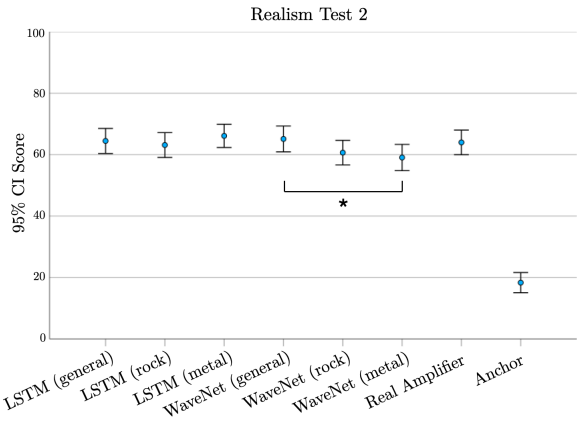

Figure 6: *The mean scores of each model with 95% confidence intervals for the second listening test in experiment 2.*

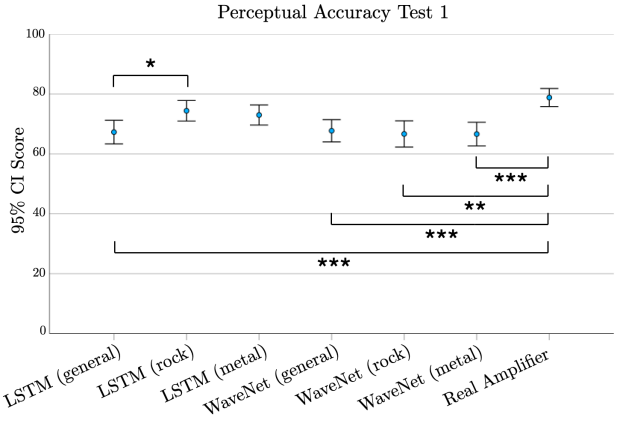

Figure 7: *Means and 95% confidence intervals for the first listening test from experiment 3: similarity to the reference. The results are averaged across all excerpts (B, D, F and H).*

returned a statistically significant difference in realism between the models ($\chi^2(6, N{=}32) = 16.162$, p = 0.013). Performing post hoc Wilcoxon signed-ranks tests showed that the *rock-specific* LSTM-RNN model was rated as significantly *more* realistic than the real amplifier (p = 0.0383, adjusted), and the *general* WaveNet-based model was also significantly *more* realistic than the real amplifier (p = 0.0234, adjusted). For higher-gain excerpts B, C, G and H, Friedman tests revealed there was no significant differences between models. Despite the Friedman test for excerpt A showing significance ($\chi^2(6, N{=}32) = 14.649$, p = 0.023), the post hoc Wilcoxon signed-ranks tests revealed no significant comparisons when considering the LSTM-RNN models alone, the WaveNet-based models alone and the 6 models versus the real amplifier.

5.6. Experiment 3 - Accuracy

The final experiment sought to evaluate the models in terms of perceptual accuracy compared to the (labelled) real amplifier.

5.6.1. Methodology

This experiment first used a MUSHRA-style test which asked subjects to rate the similarity of 7 test samples (the 6 models and a hidden reference) to a labelled reference of the real amplifier on a scale of 0 to 100. Subjects were not asked to rate one of the samples at 100. The excerpts used in this test were B, D, F and H, each on a different page, repeated once, resulting in a total of 8 pages.

An ABX test was also conducted which gave subjects a labelled reference of the real amplifier and two test samples: a hidden reference and one of the 6 models. Subjects were tasked with identifying which of the two samples was the hidden reference for excerpts A, B, C, F, G and H, randomised and repeated once, resulting in 72 trials.

5.6.2. Results

Figure 7 presents the results of the MUSHRA-style test. The hidden reference reached a mean score of just 79% (this motivated running the ABX test later). A Friedman test revealed that there was a statistically significant difference between some of the models ($\chi^2(6, N{=}168) = 49.019$, $p < 0.001$). Post hoc Wilcoxon signed-ranks tests showed that the *rock-specific* LSTM-RNN model was significantly more accurate than the *general* LSTM-RNN (p = 0.0109, adjusted). No significant differences were observed within the WaveNet-based models. The real amplifier was rated as significantly more accurate than all three WaveNet-based models as well as the *general* LSTM-RNN ($p \leq 0.001$ in each case, adjusted).

To investigate differences between excerpts, Friedman tests were run for each excerpt of the MUSHRA-style test which found that the lowest-gain excerpt F had a significant interaction ($\chi^2(6, N{=}42) = 65.812$, $p < 0.001$), while the high-gain and high-pitched excerpt H did not ($\chi^2(6, N{=}42) = 3.848$, p = 0.697). For excerpt F, post hoc Wilcoxon signed-ranks tests were run, which found that both genre-specific LSTM-RNN models were significantly more accurate than the *general* LSTM-RNN ($p < 0.001$ in both cases, adjusted). Significant differences were also found between the real amplifier versus the *general* LSTM-RNN and each genre-specific WaveNet-based model ($p < 0.001$ in each case, adjusted).

Figure 8 shows the ABX results, where the 95% and 99% critical levels are indicated (using the cumulative binomial distribution). At the 95% confidence level, it can be seen that all models could be distinguished from the reference. Using 99% confidence, however, the *rock-specific* LSTM-RNN does not exceed the critical level which suggests it was very similar to the reference. According to the cumulative binomial distribution at 95% confidence, there was no statistically significant difference between the *rock-specific* and *metal-specific* LSTM-RNN results. The *general* LSTM-RNN was identified significantly more often than both genre-specific LSTM-RNN models at $\alpha = 0.05$.

For the lowest-gain excerpts, B and F, the *rock-specific* RNN was the only model not to have been rated as significantly different to the reference at the 95% confidence level. For the high-gain and high-pitched excerpt H, however, none of the models could be distinguished from the reference.

6. DISCUSSION

In terms of preference, the mean scores for all 6 models and the real amplifier were within the -10 to 10 band (labelled “Indifferent"). There were no significant differences between the models

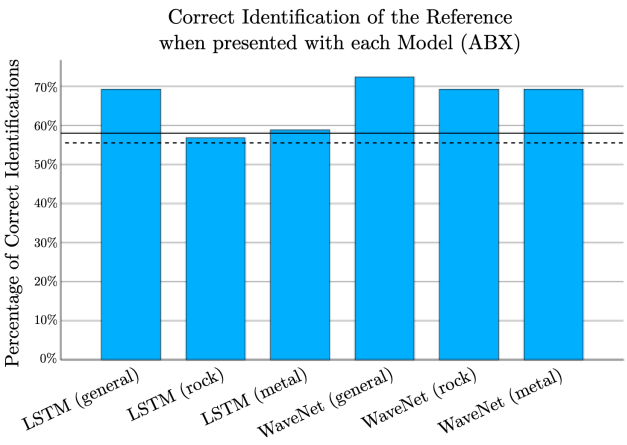

Figure 8: *The percentage of correct identifications of the reference when compared to each model in the ABX test. In this plot, lower values mean better performance. The dashed and solid horizontal lines represent the 95% and 99% critical levels, respectively.*

within either of the DNN topologies, nor were the 6 models rated as significantly different to the real amplifier. This suggests that, despite potential audible differences, the real amplifier was not preferred over any of the DNN models when accompanied with bass guitar and drums, for high-gain excerpts C, D, E and G from Table 2. Therefore, these models seem to be viable as replacements of a real guitar amplifier in a mix.

In one of the realism tests, the *metal-specific* WaveNet-based model was significantly less realistic than the real amplifier. There were no other significant differences in realism between the models and the real amplifier when considering all excerpts cumulatively. This suggests the models were generally realistic-sounding. The mean realism scores for the real amplifier were low in both tests (58% and 64%) - it is possible that this was due to subjects finding the judgement of realism difficult. The training sets did not drastically affect the realism of the models, which is most likely due to the fact that the models are already perceived as very realistic.

In terms of perceptual accuracy, the hidden reference (real amplifier) had a surprisingly low mean score of 79%, despite subjects being asked to rate the degree of similarity to the labelled reference. It is possible that this was due to not forcing subjects to rate at least one sample to 100% and/or to the reference being so close to the other samples and thus difficult to identify. The real amplifier was rated significantly higher only in comparison to the three WaveNet-based models and the *general* LSTM-RNN model in the MUSHRA-style test. This suggests that the two genre-specific LSTM-RNN models were perceived with similar accuracy to the real amplifier, which is supported by the ABX results. The *rock-specific* LSTM-RNN was not significantly distinguished from the reference in the ABX test (at $\alpha = 0.01$), where it was correctly identified only 57% of the time, suggesting it was very similar to the real amplifier. There was no significant difference between the results of the two genre-specific LSTM-RNN models, which indicates that the *metal-specific* LSTM-RNN was also perceptually close to the real amplifier.

It was found that excerpts closer to the 'edge of breakup' revealed more differences between the models. For the high-gain and high-pitched excerpt H, all models were unable to be identified from the real amplifier in the ABX test, and the MUSHRA-style accuracy test showed no significant difference between any of the models. This is supported by the realism results, where none of the models had significantly lower mean scores than the real amplifier for this excerpt, suggesting that they were all sufficiently realistic. For the lowest-gain excerpt F, however, the *rock-specific* LSTM-RNN was the only model indistinguishable from the reference, and it was rated as significantly *more* realistic than the real amplifier. The MUSHRA-style accuracy test for this excerpt revealed that both genre-specific LSTM-RNN models were more accurate than the *general* LSTM-RNN, and that the real amplifier was more accurate than three other models. The *metal-specific* WaveNet-based model was seen to be significantly less realistic than the *general* WaveNet-based model in both realism tests.

7. CONCLUSIONS AND FURTHER WORK

This paper explores the use of two popular DNN topologies for the modelling of the Peavey 6505 amplifier. Perceptual experiments were run to evaluate models trained using a general dataset versus genre-specific datasets, rated in terms of preference, realism and accuracy.

The real amplifier was not preferred to any of the DNN models when presented in a mix. The models also successfully emulated the target amplifier in terms of realism, and one of the models was even rated as *more* realistic than the real amplifier itself. Also considering that the subjects were trained listeners, these results suggest that the models can already replace the real amplifier in most music production workflows.

The genre-specific training resulted in an improvement of the performance of the LSTM-RNN topology both in terms of accuracy, and, to a lesser extent, realism. No significant difference was observed between the three training datasets for the WaveNet-based models. It is possible that this was due to WaveNet-based models being more sensitive to the choice of hyperparameters, which in this experiment were fixed for all training sets.

Results also showed that some excerpts were better than others in highlighting differences between models. More specifically, high-gain and high-pitched excerpts were perceived with the same realism and accuracy as the real amplifier for all models, while a lower-gain excerpt on the 'edge of breakup' was identified as different from the real amplifier for 5 out of 6 models. Furthermore, significant differences in accuracy between the training of the models were highlighted for this excerpt.

Future work will involve investigating whether the results generalise to different guitar amplifiers and different settings, as well as investigating the sensitivity of the individual models to hyperparameter optimisation and pruning. This may determine whether the observed effects of genre-specific training translate to optimised models, especially for a WaveNet-based CNN.

8. ACKNOWLEDGMENTS

Many thanks to all of the subjects who participated in the listening tests.

9. REFERENCES

[1] M. M. Ramírez, E. Benetos, and J. D. Reiss, "Deep learning for black-box modeling of audio effects," *Applied Sciences*, vol. 10(2), p. 638, 2020.

[2] J. Pakarinen and D. T. Yeh, "A Review of Digital Techniques for Modeling Vacuum-Tube Guitar Amplifiers," *Computer Music Journal*, vol. 33(2), pp. 85–100, 2009.

[3] F. Eichas and U. Zölzer, "Gray-Box Modeling of Guitar Amplifiers," *Journal of the Audio Engineering Society*, vol. 66(12), pp. 1006–1015, 2018.

[4] E.-P. Damskägg, L. Juvela, E. Thuillier, and V. Välimäki, "Deep Learning for Tube Amplifier Emulation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-19)*, Aalto University, Espoo, Finland, 12-17 May, 2019, pp. 471–475.

[5] A. Wright, V. Välimäki, and L. Juvela, "Adversarial Guitar Amplifier Modelling with Unpaired Data," in *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, 4-10 June, 2023, pp. 1–5.

[6] C. J. Steinmetz and J. D. Reiss, "Efficient neural networks for real-time modeling of analog dynamic range compression," in *Audio Engineering Society Convention 151*, London, UK, 2 May, 2022, pp. 1–9.

[7] A. Wright, E. P. Damskägg, L. Juvela, and V. Välimäki, "Real-time guitar amplifier emulation with deep learning," *Applied Sciences (Switzerland)*, vol. 10(3), pp. 1–18, 2020.

[8] K. Yoshimoto, H. Kuroda, D. Kitahara, and A. Hirabayashi, "Deep Neural Network Modeling of Distortion Stomp Box Using Spectral Features," in *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, Auckland, New Zealand, 7-10 December, 2020, pp. 339–345.

[9] P. Bognár, "Audio Effect Modeling with Deep Learning Methods," Ph.D. dissertation, Vienna University of Technology, 2022.

[10] M. A. M. Ramírez, "Deep Learning for Audio Effects Modelling," Ph.D. dissertation, Queen Mary University of London, 2020.

[11] Fraunhofer IDMT, *IDMT-SMT-Guitar*. Available at: https://www.idmt.fraunhofer.de/en/publications/datasets/guitar.html (Accessed: 27 February 2023), 2014.

[12] ——, *IDMT-SMT-Bass-Single-Track*. Available at: https://www.idmt.fraunhofer.de/en/publications/datasets/bass_lines.html (Accessed: 27 February 2023), 2014.

[13] A. Wright, E. P. Damskägg, and V. Välimäki, "Real-time black-box modelling with recurrent neural networks," in *Proceedings of the 22nd International Conference on Digital Audio Effects (DAFx-19)*, Birmingham, UK, 2-6 September, 2019, pp. 1–8.

[14] J. D. Parker, F. Esqueda, and A. Bergner, "Modelling of nonlinear state-space systems using a deep neural network," in *Proceedings of the International Conference on Digital Audio Effects (DAFx-19)*, Birmingham, UK, 2-6 September, 2019, pp. 165–172.

[15] E. Barbour, "The cool sound of tubes [vacuum tube musical applications]," *IEEE Spectrum*, vol. 35(8), pp. 24–35, 1998.

[16] D. Yeh, B. Bank, and M. Karjalainen, "Nonlinear modeling of a guitar loudspeaker cabinet," in *11th International Conference on Digital Audio Effects*, Espoo, Finland, 1-4 September, 2008, pp. 300–304.

[17] T. Vanhatalo, P. Legrand, M. Desainte-Catherine, P. Hanna, A. Brusco, G. Pille, and Y. Bayle, "A Review of Neural Network-Based Emulation of Guitar Amplifiers," *Applied Sciences (Switzerland)*, vol. 12(12), pp. 1–26, 2022.

[18] K. Bloemer, *GuitarML*. Available at: https://github.com/GuitarML (Accessed: 12 January 2023), 2022.

[19] T. Koker, *pedalnet*. Available at: https://github.com/teddykoker/pedalnet (Accessed: 12 January 2023), 2020.

[20] E.-P. Damskägg, *WaveNetVA*. Available at: https://github.com/damskaggep/WaveNetVA (Accessed: 20 March 2023), 2020.

[21] J. Chowdhury, "RTNeural: Fast Neural Inferencing for Real-Time Systems," *arXiv preprint*, 2021.

[22] A. Wright, *Automated-GuitarAmpModelling*. Available at: https://github.com/Alec-Wright/Automated-GuitarAmpModelling (Accessed: 12 January 2023), 2021.

[23] J. P. Herbst, "Shredding, tapping and sweeping: Effects of guitar distortion on playability and expressiveness in rock and metal solos," *Metal Music Studies*, vol. 3(2), pp. 231–250, 2017.

[24] Universal Audio, Inc., *Diezel Herbert Amplifier*. Available at: https://www.uaudio.com/uad-plugins/guitar-bass/diezel-herbert-amplifier.html (Accessed: 05 April 2023), 2023.

[25] T. D. Sunnerberg, "Analog Musical Distortion Circuits for Electric Guitars," Master's thesis, Rochester Institute of Technology, 2019.

[26] Rivera Research and Development Co., *RockCrusher™ Owner's Manual*. Available at: https://www.manualslib.com/manual/1930406/Rivera-Rockcrusher.html?page=6#manual (Accessed: 25 February 2023), 2010.

[27] Universal Audio, Inc., *Apollo Twin X Hardware Manual*. Available at: https://media.uaudio.com/support/manuals/hardware/Apollo%20Twin%20X%20Hardware%20Manual.pdf (Accessed: 22 February 2023), 2021.

[28] M. Doidic, M. Mecca, M. Ryle, C. Senffner *et al.*, "Tube Modeling Programmable Digital Guitar Amplification System," US Patent 5789689, August 1998.

[29] ITU-R, "Method for the subjective assessment of intermediate quality level of audio systems," *Recommendation ITU-R BS.1534-3*, 2015.

[30] ——, "Methods for the subjective assessment of small impairments in audio systems," *Recommendation ITU-R BS.1116-3*, 2015.

[31] C. Mendonça and S. Delikaris-Manias, "Statistical tests with MUSHRA data," in *Audio Engineering Society Convention 144*, Milan, Italy, May 23-26 2018, pp. 859–868.