Audio Engineering Society

Conference Paper 30

Presented at the AES International Conference on Machine Learning and Artificial Intelligence for Audio
2025 September 8–10, London, UK

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Perceptual Evaluation of Machine Learning and Non-ML Emulations of the Vox AC30 Amplifier

Mario Alberto Vallejo Reyes[1], Gavin Kearney[1], and Michael McLoughlin[1]

[1]*School of Physics, Engineering and Technology, University of York*

Correspondence should be addressed to Mario Alberto Vallejo Reyes (mario.vallejo@york.ac.uk)

ABSTRACT

Twenty-two listeners evaluated the perceptual similarity of three commercially available emulations of a Vox AC30 amplifier to the original hardware reference under controlled conditions: (1) a machine-learning plugin trained on the exact test amplifier, (2-3) two non-ML plugins modeling unspecified AC30 units. Using MUSHRA methodology, results showed the unit-specific ML plugin achieved perceptual indistinguishability from the hardware reference in 1/3 cases and outperformed both non-ML alternatives in 5/6 pairwise comparisons. This demonstrates that while non-ML plugins necessarily generalize across hardware units, ML techniques can achieve high perceptual fidelity when trained on target amplifiers. This represents a previously unattainable capability with significant implications for audio preservation and production.

1 Introduction

Over time, digital technology has progressively shaped guitar amplifier emulation, with increasingly sophisticated techniques emerging to capture the nuanced, nonlinear behaviors of analog amplifiers [1]. As studio environments grow more diverse and flexible [2], offering digital versatility while preserving cherished hardware characteristics has become crucial.

Recent innovations include the application of machine learning (ML) in amplifier emulation, a shift that promises new levels of authenticity compared to conventional digital signal processing (DSP). However, skepticism persists in the broader audio community: a survey of 1,500 participants by Bedroom Producers Blog [3] reflects limited enthusiasm for AI tools in music production. This intersection of technological advancement and user acceptance underscores that in the field of music production and audio engineering, accurately reproducing and quantifying sound quality remains a key focus of technological development [4]. While objective measurements, such as numerical measures of interference, noise, and artifacts, are indispensable, subjective listening tests remain critical in determining perceptual quality [5].

The present study evaluates perceptual similarities between ML-based and non-ML amplifier emulations relative to a genuine Vox AC30 amplifier through listening tests following the MUSHRA protocol (Multi-Stimulus Test with Hidden Reference and Anchor). The goal is to offer empirical data on these comparisons. While existing research has focused on experimental or prototype ML emulations, this study evaluates commercially available plugins, including a machine learning-based

model (Neural Amp Modeler [6]) and two widely used non-ML emulation tools: Waves GTR3 (Version 14.12 used in this study, released in 2022 [7]; originally released in 2007 [8]) and IK Multimedia AmpliTube 5 (Version 5.7.5 used in this study, released in 2020 [9]; originally released in 2002 [10]). These non-ML plugins remain relevant benchmarks due to their widespread use in contemporary workflows. Ultimately, the findings may contribute to ongoing discussions within the music industry regarding the role and acceptance of AI-based tools in production workflows.

1.1 Previous Findings

Pakarinen and Yeh [1] highlight that the emulation of vacuum-tube guitar amplifiers has long been a challenging endeavor due to the complex nonlinearities and dynamic behaviors inherent in their design. According to their work, traditional approaches, such as analog solid-state emulations and digital signal processing (DSP) techniques, have sought to replicate the distinctive tonal characteristics of tube amplifiers. Pakarinen and Yeh explain that these methods often rely on static waveshaping, oversampling, and circuit simulation to approximate the amplifier's response. However, they note that accurately capturing the intricate interdependencies and dynamic effects, such as grid capacitor charging, cathode bypass behavior, and power amplifier sagging, remains a significant challenge. While DSP-based emulations offer flexibility and efficiency, enabling real-time modeling and integration into software plugins, Pakarinen and Yeh emphasize that the complex dynamical nonlinearities of the tube-amplifier circuit make accurate real-time simulation extremely demanding. Building on the foundational insights provided by Pakarinen and Yeh, the discussion progresses toward exploring more advanced techniques, such as machine learning (ML), which aim to leverage data-driven approaches to model the complex behaviors of tube amplifiers more effectively.

Zhang et al. [11] investigated modeling vacuum-tube guitar amplifiers using LSTM networks, focusing on replicating tube-driven distortion and tonal characteristics. Their approach involved exploring various hyperparameter configurations and assessing model performance through objective metrics and casual feedback from semi-professional guitarists. Although their method approximated tube amplifier responses, guitarist evaluations indicated perceptual differences that still require refinement.

Wright et al. [12] demonstrated the feasibility of real-time emulation by comparing WaveNet-inspired and recurrent neural network architectures. They trained models using only a few minutes of audio data and evaluated them through subjective listening tests. The results showed that while some models still deviated perceptually from the reference signals, others were virtually indistinguishable, underscoring the promise of ML methods for efficient, perceptually convincing emulations.

Juvela et al. [13] proposed an automated, data-driven framework using an LSTM network conditioned on amplifier control parameters. This approach aimed to capture the amplifier's full control range and dynamics. Their perceptual tests, involving expert listeners, found that the proposed model achieved subjective quality comparable to that of a white-box Simulation Program with Integrated Circuit Emphasis (SPICE)[14] simulation. SPICE, an industry-standard tool for electrical circuit simulation, requires significant expert effort and computational resources, often taking hours to render audio samples. In contrast, the LSTM-based model achieved comparable perceptual quality while maintaining real-time computational efficiency.

Collectively, these studies highlight the growing sophistication of ML-based amplifier modeling. However, perceptual discrepancies remain, Zhang et al. [11] noted that guitarists still heard differences in emulated tube responses, and Wright et al. [12] found certain models deviated perceptually from reference signals. Nevertheless, these studies indicate that ML-based modeling continues to improve in realism, flexibility, and efficiency.

2 Methods

2.1 Stimuli Creation

The dataset consists of three 12-second clean guitar recordings from a Fender Stratocaster, recorded through a Universal Audio 76 interface at 48 kHz/24-bit. All performances were by the same musician using identical equipment and settings, with each sample featuring rhythm guitar playing simple pop-rock chord progressions that differ only in the specific chords and tempo. The recordings adhere to MUSHRA guidelines for stable audio content [15].

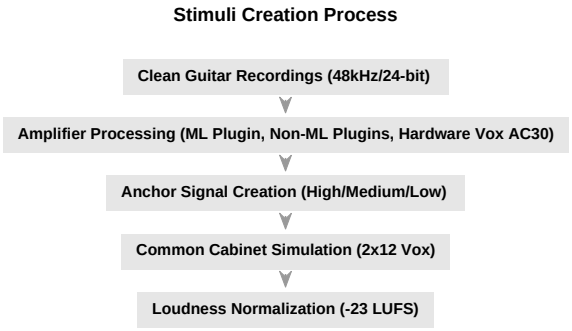


Fig. 1: Flowchart of the stimuli creation process.

Neural Amp, a machine learning plugin developed by Steve Atkinson, was used to profile a Vox AC30 amplifier. The profiling process involved playing Neural Amp's provided test signal through the amplifier's High Input Top Boost channel at 48kHz/24-bit, recorded at -18 dBFS. The signal chain included a Focusrite interface, Radial Engineering re-amp box for impedance matching, and a Torpedo Captor 16 load box to attenuate the amplifier's speaker output signal. This setup bypassed the built-in speaker, ensuring a pure recording of the amplifier's output before the cabinet. All amplifier controls were standardized at midpoints with time-based effects disabled. The recorded output was processed through Neural Amp's Google Colab environment to generate the amplifier profile (.NAM file). The final samples were created in Ableton Live 12 [16] (chosen for VST/AU compatibility) using the Neural Amp plugin with the generated profile.

IK Multimedia Amplitube and Waves GTR plugins were used to create non-machine learning samples using their Vox AC30 emulations. Both plugins were configured in Pro Tools [17] (48kHz/24-bit) with controls set to midpoints, time-based effects disabled, and cabinet simulation bypassed to match the configuration of the real amplifier recording. Neither plugin specifies whether it emulates the high gain or low gain input section of the amplifier. The clean guitar samples were processed through both emulations and exported for subsequent cabinet processing and loudness normalization.

To create the hardware Vox AC30 samples, the clean guitar tracks were processed through the amplifier using a 48kHz/24-bit Pro Tools session. The signal chain included a Focusrite interface, Radial Engineering re-amp box for impedance matching, and Torpedo Captor 16 load box for signal attenuation. Signals were routed to the amplifier's High Input Top Boost channel with all controls set to midpoints and time-based effects disabled. The processed audio was recorded back into Pro Tools and exported for subsequent cabinet processing and normalization.

2.1.1 Creating the Anchor Signals

MUSHRA methodology requires high, medium and low-quality anchor signals to validate participant responses. The high quality anchor is the same as the reference file (hardware Vox AC30). The standard medium and low quality anchors are created using 7 kHz (medium-quality) and 3.5 kHz (low-quality) high-cut filters. However, these standard cutoff frequencies produced no perceptible quality degradation in guitar amplifier evaluation due to their limited high-frequency content above 5 kHz. Therefore, the standard octave reductions from 20 kHz (1.5 and 2.5 octaves) were doubled, resulting in new cutoff frequencies of 2466 Hz for medium-quality and 616 Hz for low-quality anchors. These modified frequencies created perceptually distinct quality levels appropriate for guitar amplifier evaluation.

2.2 Stimuli Processing

To ensure fair amplifier comparisons during listening tests, all signals were recorded directly from amplifier outputs without cabinets to focus solely on amplifier characteristics. A consistent cabinet simulation—the '2_12InchOpenBack' model from Waves GTR, based on a Vox AC30 Celestion Alnico Blue—was applied to standardize output. This process involved routing all signals through the cabinet emulation in Pro Tools, exporting the processed files as WAVs, and normalizing them in MATLAB.

Normalization followed the EBU R128 standard [18], with integrated loudness and gain adjustments implemented in MATLAB using MathWorks' recommended methods [19]. All files were adjusted to -23 LUFS. Subjective evaluations by the researchers confirmed consistent loudness, reducing bias and allowing participants to focus on tonal differences among emulations.

To exemplify both the spectral differences between the hardware amplifier and its emulations and between tasks, Figure 2 shows a 1-second-excerpt spectrogram of the hardware amplifier alongside its emulations, and

Figure 3 shows 1-second-excerpt spectrograms illustrating spectral differences between tasks for the ML stimuli.

2.3 MUSHRA Listening Test

This study employed GoListen [20], a web-based platform to implement the MUSHRA listening test methodology. The test consisted of three evaluation tasks. Each task comprised seven audio files: a reference (hardware Vox AC30), the machine learning emulation, two non-machine learning emulations, and three anchor signals. The anchors included a high-quality anchor (an audio file identical to the hardware reference), as well as medium quality and low quality anchors created using high-cut filters.

Participants were provided with bilingual instructions in English and Spanish, guiding them to rate each sample's similarity to the reference on a scale from 0 (no similarity at all) to 100 (highly similar).

The platform randomized the order of stimuli for each participant to minimize potential bias from fixed sequences. Brief supplementary questions gathered information about participants' backgrounds and listening equipment.

2.4 Statistical Analysis of Perceptual Differences

Only responses that met predefined inclusion criteria were retained for analysis: participants had to rate File 04 (the identical-to-reference anchor) as the highest or second-highest, and File 05 (the low-quality anchor) as the lowest. This ensured attentiveness and allowed for cases where some emulations were perceptually indistinguishable from the reference.

Descriptive statistics (medians and interquartile ranges, 25th to 75th percentiles) were calculated to summarize central tendency and dispersion. These were visualized with violin plots to support intuitive comparison of rating distributions across conditions.

To examine perceptual differences across the three emulation types (ML, Non-ML A, Non-ML B), participants' ratings were compared using either parametric or non-parametric methods, depending on the distribution and variance characteristics of the data.

Normality of rating distributions within each emulation group was assessed using the Shapiro–Wilk test ($\alpha = 0.05$). Homogeneity of variances across groups was evaluated using Levene's test. When variances were equal (Levene's $p > 0.05$), one-way ANOVA was used for group comparison, even when one group showed mild non-normality, due to ANOVA's robustness with balanced group sizes. When both assumptions were violated, the Kruskal–Wallis test[21] was used as a non-parametric alternative.

Post-hoc comparisons were performed when significant group differences were found: Tukey's HSD followed ANOVA, and Dunn's procedure (with multiple comparison correction) followed Kruskal–Wallis. For non-significant results ($p > 0.05$), Bayesian t-tests with Cauchy priors (scale = 0.707) were conducted using raw scores for normally distributed data and rank-transformed data otherwise, to quantify evidence strength for null effects.

3 Participants

3.1 Participants Background

The study included 22 participants: 5 university lecturers in audio engineering/music technology, 7 professional sound engineers/music producers, 6 students in music technology or audio engineering, and 4 early-career individuals. Individual experience ranged from newcomers to over 30 years. Participants' areas of expertise included mixing, recording, production, research, and live sound engineering.

3.1.1 Participants Equipment

Participants completed listening tests either in-person at the University of York (AudioLab and Music Research Centre) or online. Both formats followed identical procedures, differing only in initial instruction delivery: the researcher provided an introductory briefing for in-person sessions, while online participants accessed written instructions.

In-lab tests utilized a Mac computer (MacBook Pro 13", 2022), a Focusrite Scarlett 18i20 audio interface (2nd Gen), and open-back Sennheiser HD 650 headphones. Online participants were required to use studio-grade monitoring equipment. The specific audio equipment used by participants is summarized in Table 1.

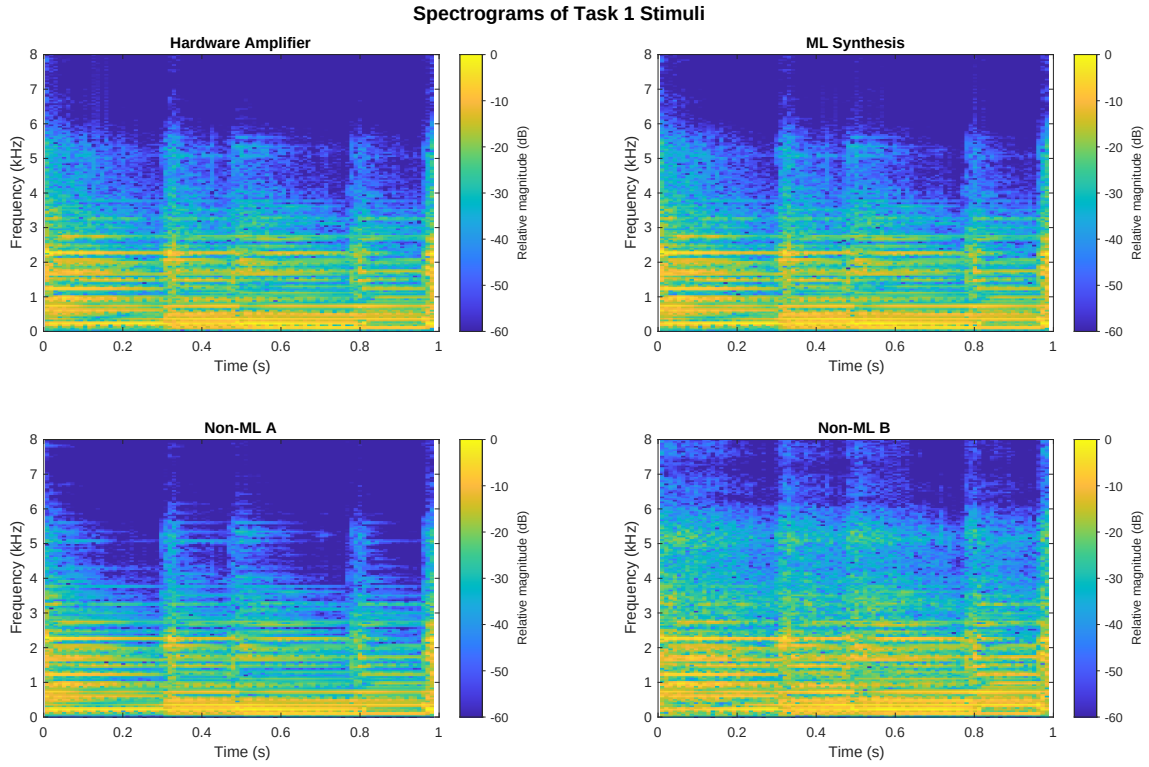


Fig. 2: 1-second spectrograms for Task 1 stimuli: Hardware Vox AC30, ML emulation, non-ML A, non-ML B.

Table 1: Summary of audio equipment used by participants

Equipment	Type
ADAM Audio S2A	Speakers
ADAM Audio T5V	Speakers
ADAM Audio T7V	Speakers
AKG K52	Headphones
beyerdynamic DT770 80OHM	Headphones
Genelec 4020C	Speakers
JBL 305P MKII	Speakers
Sennheiser HD650	Headphones
Yamaha HS8 w/ADAM Audio Subwoofer	Speakers

4 Results

Tables 2, 3, and 4 summarize the median similarity ratings and interquartile ranges (25th–75th percentiles) for Tasks 1, 2, and 3, respectively. The corresponding violin plots, which display the rating distributions (kernel-density outline) together with the median, interquartile range, and minimum/maximum values, are shown in Figures 4, 5, and 6.

Significant differences in perceived similarity among emulation conditions were observed for all three tasks. ANOVA results showed statistically significant group differences for Task 1 ($F(5,96) = 69.28$, $p < 0.001$) and Task 3 ($F(5,90) = 44.45$, $p < 0.001$). For Task 2, assumptions of normality (Shapiro-Wilk, Non-ML B: $p = 0.013$) and homogeneity of variance (Levene's test: $p = 0.015$) were violated, and the Kruskal-Wallis test was performed, which confirmed significant differences ($\chi^2(2) = 16.28$, $p < 0.001$).

Post-hoc pairwise comparisons (Tukey HSD for Tasks 1 and 3, Dunn's test for Task 2) are summarized in Tables 5, 6, and 7. These tables report the mean differences, confidence intervals, and exact p-values for each pairwise contrast.

Additionally, Table 8 presents the outcomes of Bayesian t-tests for pairwise comparisons that did not reach significance in the frequentist analysis, offering complementary evidence based on Bayes Factors.

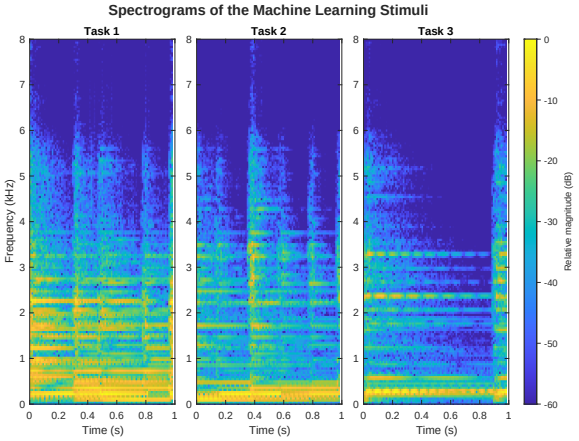


Fig. 3: 1-second-excerpt spectrograms of the machine-learning–generated guitar stimuli after amplifier and cabinet processing and loudness normalization.

Table 2: Task 1 Median Similarity Ratings and Interquartile Ranges (25th–75th percentile)

File (Condition)	Task 1
File 01 (Machine Learning)	100 (88.25–100)
File 04 (Anchor: High Quality)	96 (93.75–100)
File 02 (Non-ML A: Amplitube)	71 (62.25–80.25)
File 06 (Anchor: Medium Quality)	42 (15.75–60)
File 03 (Non-ML B: Waves)	36 (18.75–46.25)
File 05 (Anchor: Low Quality)	0 (0–20)

Table 3: Task 2 Median Similarity Ratings and Interquartile Ranges (25th–75th percentile)

File (Condition)	Task 2
File 04 (Anchor: High Quality)	99 (95.75–100)
File 03 (Non-ML B: Waves)	95 (85.25–100)
File 01 (Machine Learning)	85 (59.75–92)
File 02 (Non-ML A: Amplitube)	71 (55.75–82)
File 06 (Anchor: Medium Quality)	46 (30.75–50)
File 05 (Anchor: Low Quality)	0 (0–16.25)

5 Discussion

5.1 Interpretation of Results: Task 1

Parametric assumptions were satisfied, and a one-way ANOVA confirmed an overall effect of condition on similarity ratings, $F(5,96) = 69.28,\ p < 0.001$. The machine-learning (ML) emulation received the highest median similarity score (100), slightly above the median for the high-quality anchor (File 04, identical to

Table 4: Task 3 Median Similarity Ratings and Interquartile Ranges (25th–75th percentile)

File (Condition)	Task 3
File 04 (Anchor: High Quality)	100 (95.5–100)
File 01 (Machine Learning)	82 (66.5–90)
File 02 (Non-ML A: Amplitube)	74 (57.5–81.5)
File 06 (Anchor: Medium Quality)	48 (22–66)
File 03 (Non-ML B: Waves)	41.5 (22.5–58.5)
File 05 (Anchor: Low Quality)	0 (0–21.5)

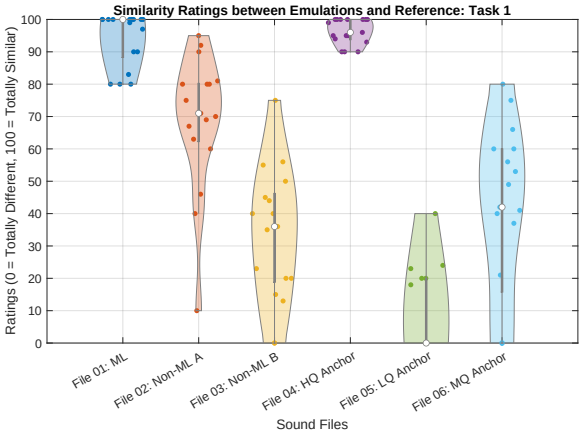


Fig. 4: Similarity ratings of emulations to the reference in Task 1.

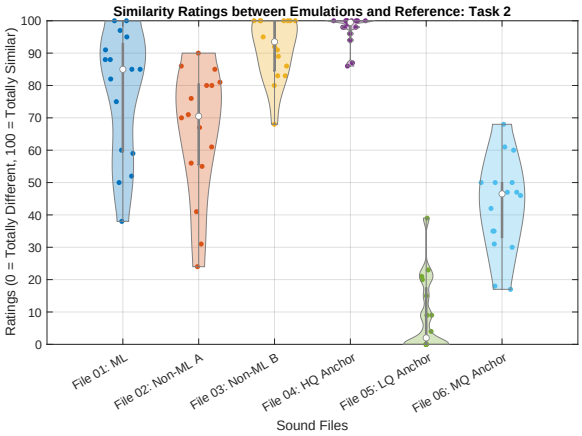


Fig. 5: Similarity ratings of emulations to the reference in Task 2.

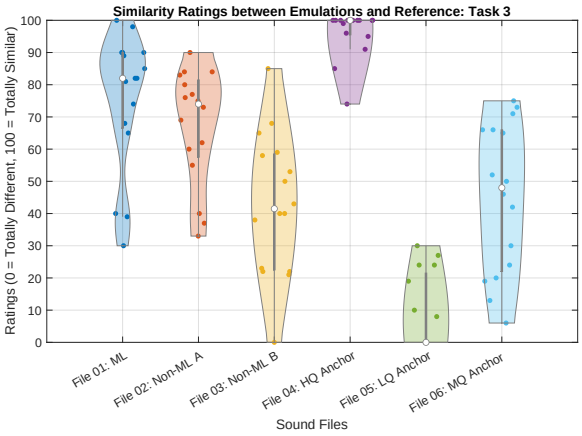


Fig. 6: Similarity ratings of emulations to the reference in Task 3.

Contrast	MeanDiff	95% CI	p-value
ML vs Non-ML A	21.33	[10.22, 32.44]	< 0.001
ML vs Non-ML B	57.67	[42.12, 73.21]	< 0.001
ML vs HQ Anchor	3.67	[-11.22, 3.89]	0.616
Non-ML A vs Non-ML B	36.33	[17.40, 55.27]	< 0.001

Table 5: Task 1: Tukey HSD pairwise contrasts ($\alpha = 0.05$).

Contrast	MeanDiff	95% CI	p-value
ML vs Non-ML A	12.69	[–16.06, 41.43]	0.962
ML vs Non-ML B	–15.78	[–44.52, 12.96]	0.820
ML vs HQ Anchor	–23.44	[–52.18, 5.31]	0.226
Non-ML A vs Non-ML B	–28.47	[–57.29, 0.35]	0.055

Table 6: Task 2: Dunn–Sidak post hoc pairwise comparisons ($\alpha = 0.05$).

Contrast	MeanDiff	95% CI	p-value
ML vs Non-ML A	7.81	[–0.94, 16.57]	0.094
ML vs Non-ML B	32.25	[8.79, 55.71]	0.005
ML vs HQ Anchor	–21.06	[–38.35, –3.77]	0.013
Non-ML A vs Non-ML B	24.44	[4.27, 44.61]	0.014

Table 7: Task 3: Tukey HSD pairwise contrasts ($\alpha = 0.05$).

the hardware Vox AC30) (96). Post-hoc Tukey tests, however, detected no statistically reliable difference between these two conditions ($p = 0.616$, $95\% CI = [-11.22, 3.89]$). This outcome indicates that listeners treated the ML emulation and the high-quality anchor as perceptually indistinguishable; the small numerical advantage for ML is consistent with random rating

Table 8: Bayesian evidence for selected pairwise comparisons ($p > 0.05$)

Comparison	(BF_{10})	Interpretation
Task 2: ML vs HQ Anchor	>100	Decisive evidence for Anchor superiority
Task 2: ML vs Non-ML A	2.98	Anecdotal evidence for ML superiority
Task 2: ML vs Non-ML B	2.00	Anecdotal evidence for Non-ML B superiority
Task 3: ML vs Non-ML A	5.05	Substantial evidence for ML superiority

variability when two stimuli are nearly identical. In contrast, ML received statistically significantly higher similarity ratings than non-ML A (median 71) and non-ML B (median 36), each $p < 0.001$. Overall, in Task 1 the ML emulation was perceptually indistinguishable from the high-quality anchor while exhibiting statistically significantly higher similarity ratings than either non-ML emulation.

5.2 Interpretation of Results: Task 2

The Kruskal-Wallis test revealed significant differences among conditions ($\chi^2(2) = 16.28$, $p < 0.001$). Dunn-Sidak post-hoc tests showed the ML emulation (median = 85) was not significantly different from the high-quality anchor (median = 99; $p = 0.226$), though Bayesian analysis of ranks decisively supported the anchor's superiority. ML showed no significant difference from Non-ML A (medians = 85 vs 71; $p = 0.962$), with Bayesian results suggesting a weak trend favoring ML. For ML versus Non-ML B (medians = 85 vs 95; $p = 0.820$), no significant difference was observed, but Bayesian analysis indicated a weak trend favoring Non-ML B.

Overall, in Task 2 ML was distinguishable from the hardware reference through Bayesian analysis, non-significantly outperformed Non-ML A but with Bayesian support, and showed weak Bayesian evidence for inferiority to Non-ML B.

5.3 Interpretation of Results: Task 3

Parametric assumptions were met, and a one-way ANOVA yielded $F(5,90) = 44.45$, $p < 0.001$. Tukey tests indicated that, in terms of similarity to the hardware Vox AC30, ML was statistically significantly less similar than the high-quality anchor, with medians of 82 and 100, respectively ($p = 0.013$). ML was statistically significantly more similar than non-ML B ($p = 0.005$; medians 82 vs 41.5). The difference between ML and non-ML A (medians 82 vs 74) did

not reach significance ($p = 0.094$); a Bayes factor of $BF_{10} = 5.05$ nevertheless provided substantial evidence favouring an ML advantage. Overall, in Task 3 ML was distinguishable from the hardware amplifier, statistically significantly outperformed non-ML B, and showed a Bayesian-supported but non-significant advantage over non-ML A.

5.4 Across-Excerpt Synthesis

Comparing ML to the high-quality anchor, ML was indistinguishable in one of three tasks and distinguishable in the other two. In Task 2, the high-quality anchor's higher similarity ratings were not statistically significant, but Bayesian analysis decisively supported this advantage. In Task 3, the high-quality anchor's similarity ratings were significantly higher. Across the six ML–non-ML pairwise similarity comparisons to the hardware Vox AC30, ML outperformed in five cases: three were statistically significant in favor of ML, two were non-significant but had Bayesian support for ML, and one was non-significant but Bayesian-supported in favor of the non-ML system (the ML B comparison).

It should be noted that the ML model was profiled on the specific Vox AC30 unit used as the reference in the listening test, whereas the non-ML emulations were modeled after different units of the same amplifier model. This introduced an advantage for the ML condition, though its high similarity to the profiled hardware still provides a valid basis for assessing perceptual fidelity.

5.5 Listener Expertise and Perceptual Ranking Trends

To assess whether listener expertise influenced rating trends, mean similarity scores were analyzed separately for experts and non-experts whose responses met the validity criteria. Although no normality or post-hoc tests were applied due to small group sizes, clear patterns emerged. In Task 1, both groups identified the same top three conditions, but their order differed: experts ranked ML first, followed by the reference and Non-ML A; non-experts placed the reference first, followed by ML and Non-ML A. In Tasks 2 and 3, both groups showed identical rankings. This convergence suggests that listener experience had minimal effect on overall perceptual trends.

6 Conclusion

This study evaluated the perceptual similarity of a commercially available machine-learning (ML) guitar amplifier emulator against two non-ML alternatives and three quality anchors, using the hardware Vox AC30 as reference.

In Task 1, ML was statistically indistinguishable from the high-quality anchor and significantly closer to the reference than both non-ML alternatives. In Task 2, ML was rated less similar than the high-quality anchor; the gap was non-significant, but Bayesian analysis favored the anchor. ML held a non-significant, Bayes-supported edge over non-ML A, while the ML-B comparison was non-significant with Bayesian support for the non-ML system. In Task 3, ML was again significantly less similar than the high-quality anchor, significantly outperformed non-ML B, and showed a non-significant, Bayes-supported advantage over non-ML A. Overall, ML exceeded non-ML emulations in five of six pairwise comparisons, either significantly or with Bayesian support.

Task 2 failed the Shapiro-Wilk normality test for the Non-ML B condition and Levene's test for homogeneity of variance across the three emulation groups. The emulation rating distributions for Task 2 exhibit pronounced skewness, indicating a non-normal distribution of responses. This skewness may arise from the relatively small number of listeners or from intrinsic characteristics of the audio stimulus, such as its tempo, dynamic range, or tonal content. Studies with larger cohorts and a wider variety of clips will help determine which factor predominates.

These results extend prior research—such as Juvela et al.,[13], Wright et al.,[12], and Zhang et al.,[11]—by demonstrating that modern ML-based emulations, when profiled on the test amplifier, can achieve perceptual indistinguishability from the reference unit and generally outperform traditional non-ML approaches modeling different instances of the same amplifier design.

In practical terms, ML-based emulation emerges as a robust option for applications requiring high perceptual fidelity, although its performance still varies with programme material. Future work should include more musical content and apply this evaluation approach to other emulated devices, such as microphones or recording chains, to test generalizability.

References

[1] Pakarinen, J. and Yeh, D. T., "A Review of Digital Techniques for Modeling Vacuum-Tube Guitar Amplifiers," *Computer Music Journal*, 33(2), pp. 85–100, 2009, doi:10.1162/comj.2009.33.2.85.

[2] "The End of the World as We Know It: The Changing Role of the Studio in the Age of the Internet," in *The Art of Record Production*, pp. 98–112, Routledge, 2012, ISBN 9781409405627.

[3] Bedroom Producers Blog, "AI Music Survey: How 1,500 Music Producers Use AI For Music Production," https://bedroomproducersblog.com/2023/05/30/ai-music-survey, 2023, accessed: 2024-05-24.

[4] Bech, S. and Zacharov, N., *Perceptual Audio Evaluation: Theory, Method and Application*, Wiley, 1. aufl. edition, 2007, ISBN 9780470869239.

[5] Kornycky, J., Gunel, B., and Kondoz, A., "Comparison of Subjective and Objective Evaluation Methods for Audio Source Separation," *The Journal of the Acoustical Society of America*, 123(5 Supplement), pp. 3569–3569, 2008.

[6] Atkinson, S., "Neural Amp Modeler," https://www.neuralampmodeler.com, 2024, accessed: 2024-03-12.

[7] Waves Audio, "Waves V14 Release Notes," https://www.waves.com/downloads/release-notesv14, 2022, accessed: 2025-07-03.

[8] MusicRadar, "New amps in Waves GTR3," https://www.musicradar.com/news/guitars/new-amps-in-waves-gtr3-33152, 2007, accessed: 2025-07-03.

[9] MusicRadar, "It's lift-off for AmpliTube 5 as IK Multimedia unveils the next edition of its flagship guitar-software package," https://www.musicradar.com/news/its-lift-off-for-amplitube-5-as-ik-multimedia-unveils-the-next-edition-of-its-flagship-guitar-software-package, 2020, accessed: 2025-07-03.

[10] IK Multimedia, "About Us — Company Timeline," https://www.ikmultimedia.com/about-us/, 2002, accessed: 2025-07-03.

[11] Zhang, Z., Olbrych, E., Bruchalski, J., McCormick, T. J., and Livingston, D. L., "A Vacuum-Tube Guitar Amplifier Model Using Long/Short-Term Memory Networks," in *SoutheastCon 2018*, pp. 1–5, IEEE, 2018, ISBN 1538661330.

[12] Wright, A., Damskägg, E.-P., Juvela, L., and Välimäki, V., "Real-Time Guitar Amplifier Emulation with Deep Learning," *Applied Sciences*, 10(3), pp. 766–, 2020.

[13] Juvela, L., Damskagg, E.-P., Peussa, A., Makinen, J., Sherson, T., Mimilakis, S. I., Rauhanen, K., and Gotsopoulos, A., "End-to-End Amp Modeling: From Data to Controllable Guitar Amplifier Models," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023, ISBN 1728163277.

[14] Nagel, L. W. and Pederson, D. O., "SPICE (Simulation Program with Integrated Circuit Emphasis)," Technical Report UCB/ERL M382, University of California, Berkeley, EECS Department, 1973.

[15] International Telecommunication Union, "Recommendation ITU-R BS.1534-1: Method for the subjective assessment of intermediate quality level of coding systems," Technical report, International Telecommunication Union, 2003, accessed: 2024-03-12.

[16] Ableton, "Ableton Live," https://www.ableton.com/en/live, 2025, accessed: 2025-03-13.

[17] Avid Technology, "Pro Tools," https://www.avid.com/pro-tools, 2025, accessed: 2025-03-13.

[18] EBU, "128, Loudness Normalisation and Permitted Maximum Level of Audio Signals," *EBU Recommendation, Geneva*, 2014.

[19] MathWorks, "Loudness Normalization in Accordance with EBU R 128 Standard," https://uk.mathworks.com/help/audio/ug/loudness-normalization-in-accordance-with-ebu-r-128-standard.html, 2024, accessed: 2024-03-12.

[20] Barry, D., Zhang, Q., Sun, P. W., and Hines, A., "Go Listen: An End-to-End Online Listening Test Platform," *Journal of Open Research Software*, 9(1), pp. 20–, 2021.

[21] Kruskal, W. H. and Wallis, W. A., "Use of Ranks in One-Criterion Variance Analysis," *Journal of the American Statistical Association*, 47(260), pp. 583–621, 1952.