DOI: 10.71568/daga2026.154

Hybrid Modeling of Guitar Cabinets Using FIR Filters and Neural Networks: Perceptual and Technical Evaluation

Simon Grimm
d&b audiotechnik GmbH & Co. KG, 71522 Backnang, Deutschland, Email: simon.grimm@dbaudio.com

Abstract
Accurate modeling of guitar loudspeakers is crucial for realistic cabinet simulation in digital audio applications. While linear impulse response models are widely used, they fail to account for nonlinear effects. This study investigates the perceptual and technical relevance of nonlinear modeling using two exemplary guitar cabinet types: a closed 4×12" and an open 1×12", each with distinct electroacoustic characteristics. Hybrid models combining FIR filters and neural networks are trained to capture nonlinear behavior at realistic playback levels. Additionally, the systematic reduction of FIR filter length in linear modeling is examined for both cabinet types to determine the number of filter taps required for adequate modeling. The objective evaluation is based on the signal-to-error ratio (SER) and the Mel log-spectral distance (Mel-LSD), while perceptual assessment relies on a recently published MOS estimation method and ABX listening tests. The results show that neural networks can improve the SER, especially for the open 1×12" cabinet. However, the perceptual gains are small, and for most scenarios, a linear modeling approach is sufficient.

Introduction
Guitar loudspeakers and their sonic fingerprint have a major impact in shaping the final sound of electric guitar recordings and live performances. Unlike conventional Hi-Fi or PA loudspeakers, they are strongly colored and band-limited. They therefore contribute significantly to the instrument's character. Consequently, accurate modeling of guitar cabinets is essential for digital amp simulations, plug-in development and real-time audio processing. While linear convolution using measured impulse responses (IRs) has become a standard technique for cabinet simulation, it inherently neglects nonlinear behaviors such as cone distortion, magnetic hysteresis and mechanical compression. Previous studies have proposed nonlinear models to address these effects, including Volterra-based approaches [1] and perceptual analyses of distortion audibility [2]. More recently, machine learning techniques have been explored for loudspeaker modeling and control, offering flexible alternatives to traditional physical models [3]. This study aims to systematically evaluate whether nonlinear modeling provides perceptual benefits over linear FIR-based approaches. Two cabinet types with distinct acoustic characteristics are investigated, and four modeling strategies using neural networks in combination with FIR filters are explored. In addition, we analyze how FIR filter length truncation affects perceptual transparency in linear modeling. The goal of this work is to provide practical insights into when nonlinear modeling is necessary and how perceptually transparent cabinet simulations can be achieved using modern machine learning techniques in combination with FIR filtering.

This paper is structured as follows: First, we introduce the signal notation and model the guitar loudspeaker and microphone system using hybrid approaches of linear filtering and deep neural networks (DNNs). Subsequently, we detail the experimental setup and evaluation metrics, before presenting the results of our technical and perceptual assessments. Finally, the paper concludes with a summary of our findings.

Signal Model and DNN Variants
In this section, the signal model and corresponding notation as well as the different DNN variants are derived. The system under test (DUT) is depicted in Figure 1, comprising one or more guitar loudspeakers in an enclosure and a microphone placed in the near-field of the cabinet.

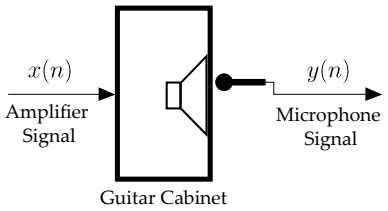

Figure 1: The guitar cabinet and microphone system

We focus solely on modeling the guitar cabinet and microphone response. The input signal $x(n)$, where n denotes the discrete-time index, is assumed to already include the guitar and amplifier signal and the sonic footprint of the amplifier-cabinet interaction, common with low-damping-factor tube amplifiers.

The output of the DUT is modeled as

$$y(n) = \hat{y}(n) + e(n), \tag{1}$$

where $\hat{y}(n)$ represents the model estimate, and $e(n)$ is the residual modeling error. In the most general case, $\hat{y}(n)$ is a functional mapping of the past N input samples:

$$\hat{y}(n) = f\left(x(n), x(n-1), \ldots, x(n-N+1)\right). \tag{2}$$

If the system is modeled as linear and time-invariant, the output estimate $\hat{y}_{lin}(n)$ is described by the convolution of the input signal $x(n)$ with the impulse response $h(n)$:

$$\hat{y}_{lin}(n) = \sum_{k=0}^{N-1} h(k) \cdot x(n-k). \tag{3}$$

Depending on the modeling variants, which are described in the following, $\hat{y}(n)$ is obtained by a variation of cascaded as well as parallel structures of the input signal, the linearly modeled signal, and the output signal of a trained neural network. The corresponding block diagram is depicted in Figure 2. The time index n is omitted for brevity.

Variant 1a)
For this basic variant, the signal x is used as the input of a neural network to produce the estimate

$$\hat{y}_{1a} = f_{DNN}(x). \tag{4}$$

Variant 1b)
This variant is similar to Variant 1a), but uses $\hat{y}_{lin}$ of Eq. 3 as the input of the neural network, to obtain

$$\hat{y}_{1b} = f_{DNN}(\hat{y}_{lin}). \tag{5}$$

Variant 2a)
In this configuration, the final estimate $\hat{y}_{2a}$ results from adding the linear system output $\hat{y}_{lin}$ to the neural network output $\hat{y}_{DNN}$. Both the linear model and the neural network use the same input signal x. This approach aims to use the neural network to model the residual signal that the linear impulse response fails to predict.

$$\hat{y}_{2a} = f_{DNN}(x) + \hat{y}_{lin}. \tag{6}$$

Variant 2b)
This variant is similar to Variant 2a), but uses the linearly modeled signal $\hat{y}_{lin}$ as the input to the neural network to generate the final estimate

$$\hat{y}_{2b} = f_{DNN}(\hat{y}_{lin}) + \hat{y}_{lin}. \tag{7}$$

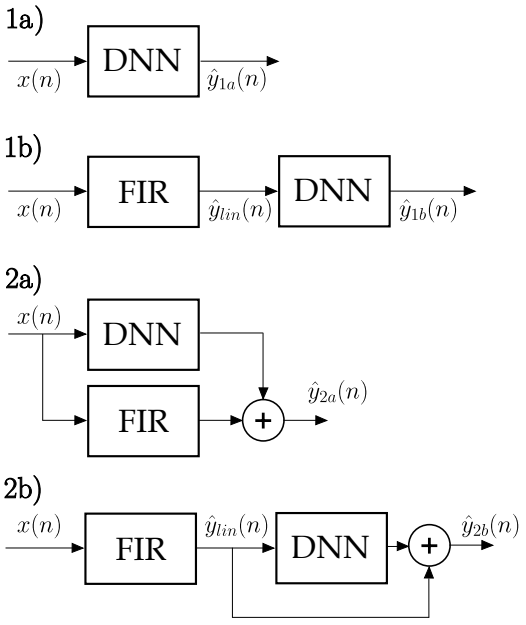

Figure 2: Modeling variants using FIR filters and DNNs

DNN Architecture

The open-source *Neural Amp Modeler* (NAM) framework [4] is used as the DNN architecture in this study. This framework implements a modified, non-autoregressive and feed-forward variant of the WaveNet architecture, specifically optimized for real-time audio applications through causal convolutions. We use the standard configuration, which comprises two stacks, each consisting of 10 dilated convolutional layers. To achieve a sufficiently large receptive field, the dilation rates increase exponentially ($d \in \{2^0, 2^1, \ldots, 2^9\}$). Each layer uses a hyperbolic-tangent-based activation function. For each modeling variant, the networks were trained over 1,000 epochs with an initial learning rate of 0.004 and a decay rate of 0.007. Training the model was managed using the Tone3000 platform [5].

Experimental Setup

For the experimental setup, two distinct guitar loudspeakers were examined: a closed-back 4×12" Mesa Boogie Rectifier Traditional cabinet, loaded with Vintage Celestion V30 speakers (8 Ω), and a 1×12" open-back Peavey Classic 30 amplifier, where only the loudspeaker component (a 16 Ω Peavey Blue Marvel speaker) was part of the setup. For both cabinets, the signals were captured with a Shure SM57 microphone. A d&b audiotechnik D90 amplifier was used to drive the loudspeakers. Measurements of the DUT were taken in a (semi-)anechoic chamber to eliminate room-acoustic influences, which are not part of the cabinet-microphone response and vary across different acoustic scenarios. The setup can be seen in Figure 3.

Figure 3: Measurement of the DUT in the (semi-)anechoic chamber

For a fair comparison, both cabinets were level-matched to produce a sound pressure level (SPL) of 112 dB(A) at one meter distance. This is regarded as a realistic stage volume and reflects the modeling of the system response at practical playing levels.

The signals provided to the DUT for training, validation, and testing (a sequence of 6 minutes) consisted of pink noise, sine and multisine tones, sine sweeps, and guitar signals of various music styles (which already include the sonic footprint of suitable guitar amplifiers, excluding the cabinet). The guitar signals were only used to compare the trained model results with the real microphone signals and were excluded from training. All signals were recorded at a sampling rate of $f_s = 48$ kHz.

Evaluation Metrics

In the following, the applied metrics for technical and perceptual evaluation are presented.

Signal-to-error ratio (SER)

The signal-to-error ratio is calculated using the time-domain signals of the measured microphone signal $y(n)$ and the error signal $e(n)$ over a sequence of M samples by

$$\mathrm{SER} = 10\log_{10}\left(\frac{\sum_{n=0}^{M-1} y(n)^2}{\sum_{n=0}^{M-1} e(n)^2}\right). \tag{8}$$

Mel log-spectral distance (Mel-LSD)

The Mel-LSD is calculated using the Mel power spectral densities P of the microphone signal and the estimated microphone signal:

$$\text{Mel-LSD} = \sqrt{\frac{1}{N_\nu}\sum_{\nu=1}^{N_\nu}\left(\log P_y(\nu) - \log P_{\hat{y}}(\nu)\right)^2} \tag{9}$$

where ν denotes the Mel frequency coefficients. For this evaluation, $N_\nu = 128$ coefficients were used, and log denotes the natural logarithm.

Zimtohrli Mean Opinion Score (MOS)

For the perceptual evaluation, we use the recently introduced "Zimtohrli" mean opinion score, which correlates quite well with music-content audio signals [6].

ABX Listening Test

For evaluation of just-noticeable difference (JND), ABX listening tests were performed. A small software app was written to quickly load the original microphone signal (A) as well as the signal to be evaluated (B). The user can then play back signal (X), which is randomly chosen from A or B and must decide which of the provided signals is X. The test was done with 20 listeners performing 10 trials per condition. Proportions correct were tested against chance (50%) using a one-sided binomial test.

Results

In the following, we first present the results for the linear modeling approach, followed by the DNN modeling results for both guitar cabinets. The accuracy is examined for both the technical and the perceptual measures.

Linear Modeling with FIR Filter

The linear modeling using an impulse response is evaluated as follows: in order to evaluate the required filter length for modeling (at $f_s = 48\,\mathrm{kHz}$), the impulse response of the dedicated cabinet is truncated and the SER and the MOS for a microphone signal and its modeled output over a sequence of several music styles are calculated. The results for both cabinet types can be seen in Figure 4 and 5. The SER increases as more filter taps are used for both cabinet types. By using about 800 filter taps for the 1×12" cabinet, an upper bound is reached, which suggests that nonlinear terms (which cannot be modeled with this approach) start to dominate the SER for this FIR filter length. The filter length of the 4×12" can be increased up to 4000 taps while still yielding an SER gain. In general, the 4×12" cabinet achieves an SER value up to 12 dB higher than the 1×12" cabinet, suggesting linear modeling is far more suitable for the 4×12" cabinet at the same SPL.

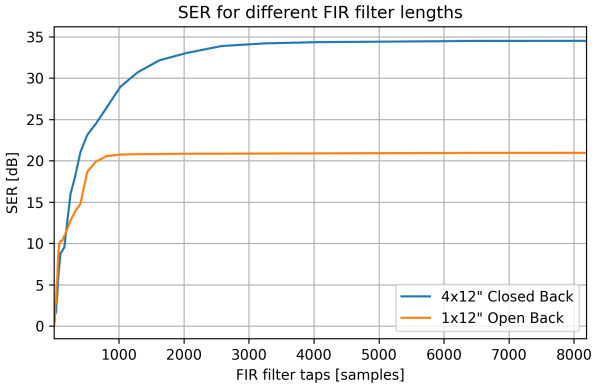

Figure 4: SER vs. filter length

The Zimtohrli MOS, on the other hand, shows quite similar results for both cabinet types if the filter length is increased. At about 600 filter taps, an upper bound on the MOS scale is reached, which cannot be further improved by increasing the filter length and therefore suggests a perceptual saturation.

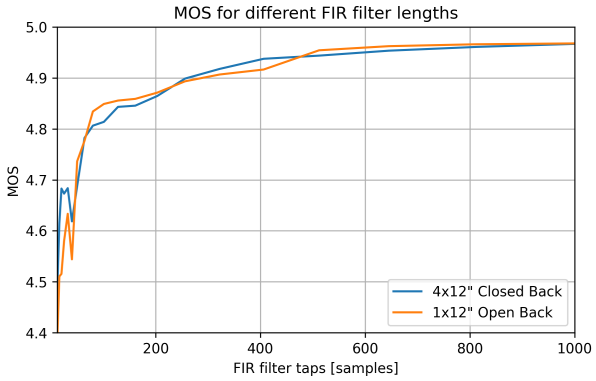

Figure 5: MOS vs. filter length

Additionally, an ABX listening test was performed as described in the previous section. A metal riff was chosen for comparison, since the error signal between linear modeling and the original microphone signal was the highest among all musical styles. A comparison between the original microphone signal of both cabinets and signals generated with FIR filtering using three different truncation lengths of the impulse response was performed. The ABX test results as depicted in Table 1 suggest that for both cabinet types, modeling the DUT with 200 filter taps can be clearly distinguished from the original signal. For a model length of 400 filter taps, the 4×12" seems to be nearly indistinguishable, while the 1×12" cabinet could be distinguished more precisely. For a modeling length of 600 filter taps, the perceptual differences become very small for both cabinet types, with the 4×12" cabinet appearing indistinguishable and the 1×12" cabinet showing only a marginally significant result. The ABX test results are supported by the corresponding Zimtohrli MOS.

Table 1: ABX listening test for linear modeling using different filter lengths (OB: open-back, CB: closed-back)

Cabinet - Filter taps	% correct (p-value)	MOS	SER (in dB)
1×12" OB - 200	86 % ($p < .001$)	4.808	9.7
1×12" OB - 400	68.5 % ($p < .001$)	4.870	12.6
1×12" OB - 600	58 % ($p = .014$)	4.920	16.0
4×12" CB - 200	82 % ($p < .001$)	4.882	10.6
4×12" CB - 400	54.5 % ($p = .115$)	4.960	19.5
4×12" CB - 600	55.5 % ($p = .069$)	4.974	23.6

Nonlinear Modeling with the DNN

In order to evaluate whether modeling of the mostly nonlinear residual further improves the accuracy, we evaluate the results for our DNN modeling variants. The results calculated over signals of several music styles are shown in Table 2.

For both cabinet types, DNN modeling mostly improves the SER compared with linear modeling (8192 filter taps, used for all following comparisons), especially for the 1×12" cabinet. Similar observations are made for the Mel log-spectral distance, which improves for all DNN modeling variants. However, the MOS values remain similar for all modeling approaches, suggesting barely noticeable differences if compared over all available music styles.

Table 2: Results for DNN modeling (SER, Mel-LSD and MOS; OB: open-back, CB: closed-back)

Cabinet	Modeling Variant				
	lin.	1a	1b	2a	2b
SER results (in dB)					
4×12" CB	34.5	33.6	35.8	39.7	35.8
1×12" OB	21.0	28.4	26.4	27.9	26.2
Mel-LSD results (in dB)					
4×12" CB	1.13	0.22	0.12	0.14	0.18
1×12" OB	0.94	0.80	0.14	0.26	0.18
MOS results					
4×12" CB	4.98	4.97	4.98	4.98	4.97
1×12" OB	4.97	4.87	4.97	4.95	4.97

To further evaluate the perceptual differences, an ABX listening comparison was made using a palm mute section of a metal riff, since this playing style exhibits the highest temporal SER among all test signals and therefore represents the worst-case scenario. We therefore evaluate whether DNN modeling helps to improve this specific signal perceptually. For the DNN modeling, Variant 2a was chosen for all listening tests. The results are shown in Table 3.

This ABX test reveals that modeling this palm mute section using a neural network can clearly improve the perceptual modeling accuracy for the 1×12" cabinet, despite not being totally indistinguishable. However, for the 4×12" cabinet, the results indicate that linear modeling seems to be sufficient. This is also supported by the corresponding Zimtohrli mean opinion score results.

Table 3: ABX listening test – Linear and DNN modeling

Modeling Variant	% correct (p-value)	MOS	SER (in dB)
1×12" - Linear	78 % ($p < .001$)	4.770	11.7
1×12" - DNN	57.5 % ($p = .020$)	4.940	19.8
4×12" - Linear	45.5 % ($p = .911$)	4.993	30.2
4×12" - DNN	50.5 % ($p = .472$)	4.996	42.0

Conclusions

Our results indicate that linear guitar cabinet modeling using FIR filters seems to be perceptually adequate for various playing styles and applications, especially for the closed 4×12" version, if the filter length is chosen to be sufficiently long. Modeling the nonlinearities using neural networks improves the SER—particularly for the open-back 1×12" version—and also reduces the Mel-LSD, but the perceptual gains remain small in practice. This suggests that, under typical playing conditions and across various musical styles, the nonlinear behavior of the loudspeaker system does not significantly contribute to the perceived sound, despite being distinguishable in some cases.

References

[1] Yeh, D. T.; Bank, B.; Karjalainen, M.: Nonlinear modeling of a guitar loudspeaker cabinet. In: Proceedings of the 11th Int. Conf. on Digital Audio Effects (DAFx-08), Espoo, Finland, 2008, pp. 89–96.

[2] Klippel, W.: Physical and perceptual evaluation of electric guitar loudspeakers. In: ICASSP 2022 – IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Singapore, 2022, pp. 341–345.

[3] Brunet, P.; Li, Y.; Kochendörfer, F.; Kubota, G.: New Applications of AI Techniques for Nonlinear Control of Loudspeakers. In: Proceedings of the Int. Congress on Acoustics, Gyeongju, Korea, 2022.

[4] Neural Amp Modeler – Open-source framework for guitar amplifier and pedal modeling, URL: `https://www.neuralampmodeler.com`

[5] Tone3000 – Community platform for NAM training and tone sharing, URL: `https://www.tone3000.com`

[6] Alakuijala, J. et al.: Zimtohrli: An Efficient Psychoacoustic Audio Similarity Metric. arXiv preprint arXiv:2509.26133, 2025. URL: `https://arxiv.org/abs/2509.26133`