



Audio Engineering Society Conference Paper 12

Presented at the AES International Conference on Artificial
Intelligence and Machine Learning in 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.

Compressing Neural Network Models of Audio Distortion Effects Using Knowledge Distillation Techniques

Riccardo Simionato¹ and Aleksander Tidemann¹

¹University of Oslo, Department of Musicology

Correspondence should be addressed to Aleksander Tidemann (aleksander.tidemann@imv.uio.no)

ABSTRACT

Neural networks have proven to be effective for modeling analog audio effects using a black-box approach. However, few can guarantee lightweight solutions suitable for real-time environments where the models must run concurrently on consumer-grade equipment. This paper explores knowledge distillation techniques for compressing recurrent neural network models for audio distortion effects, aiming to produce computationally efficient solutions with high accuracy that maintain compact model size. In particular, we consider an audio-to-audio LSTM architecture for regression tasks where small networks are trained to mimic the internal representations of larger networks, known as feature-based knowledge distillation. The evaluation was conducted on three different audio distortion effect datasets with experiments on both parametric and non-parametric data. The results show that distilled models are more accurate than non-distilled models with equal parameter size, especially for models that exhibit higher error rates. Furthermore, we observe that smaller complexity gaps between student and teacher models yield greater improvements in non-parametric cases.

1 Introduction

Through their intricate physical circuits and non-linear characteristics, analog audio effects have a unique influence on sound. Attempting to capture and digitally emulate analog audio effects is a challenging task in the field of Virtual Analog (VA) modeling [1]. VA modeling approaches are often categorized as "white-box" or "black-box", where white-box modeling [2, 3] concerns the simulations of the circuitry starting from their mathematical description, and black-box approaches [4, 5] focus primarily on analyzing the target system's measurements in order to establish mapping relationships between inputs and outputs. Using neural networks for black-box modeling of audio effects has been fairly

successful. In particular, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have shown significant promise for emulating various analog effects, such as tube amplification [6, 7], distortion [8, 9], optical dynamic range compression [10, 11], and other time-based effects [12, 13].

However, neural networks can become quite large and require significant operations per sample. This can hinder the real-time capabilities of models that strive for efficiency. In these cases, building lighter and smaller networks can be desirable. However, reducing the number of computational operations in a network is not trivial and is usually detrimental to model accuracy.

Many model compression techniques can be employed

to reduce the size of neural networks [14], including pruning, quantization, low-rank factorization, and knowledge distillation. The pruning technique focuses on diminishing the number of parameters in a neural network by identifying and removing redundant parameters that contribute less during training. This technique has also been used for VA applications [15], where authors were able to compress recurrent networks while maintaining significant performance. Quantization compresses the network by lowering the number of bits required to represent each weight. Low-rank factorization helps identify redundant parameters by utilizing matrix and tensor decomposition, which entails decomposing larger matrices into smaller ones. Lastly, knowledge distillation involves training a small network to mimic the complex representations of a larger and more complex model. Once the larger network achieves good generalization and performance on unseen data, its knowledge can be transferred to the smaller networks through distillation. This training process can maintain model performance and reduce model complexity to a significant degree [16, 17].

Knowledge distillation for audio has been primarily explored for classification tasks. Data-free distillation for audio classification was proposed in [18], using synthetic data to generate transfer sets to train student networks without the need for the original dataset. Experiments to increase the classification accuracy of student networks through increased temporal dependencies have also been examined in [19], where attention weights are extracted from transformer-based teacher models and aligned with student models using KL divergence loss. Collaborative learning frameworks with multiple audio feature representations have also been successful in improving the accuracy of networks for classification tasks [20].

However, there are only a few cases where knowledge distillation is explored for audio-related regression tasks [21] and no such study, to our knowledge, in the field of VA modeling of analog audio effects. This paper presents a preliminary study into the effects of knowledge distillation on neural networks for black-box modeling of audio distortion effects. In particular, we explore to what degree feature-based knowledge distillation can reduce the size of said networks while maintaining comparable performance, enhancing the accuracy of smaller networks. Three datasets of audio distortion effects were used for evaluation, including the Electro Harmonix Big Muff Pedal and Blackstar

HT-1 amp previously explored in [9, 15], as well as parametric and non-parametric versions of the analog modeled overdrive plugin DrDrive¹ [22]. Results compare the performance of distilled models, which are trained to mimic the behavior of the larger model, against non-distilled equivalent models, which are traditionally trained, quantitatively and through MUSHRA-style listening tests.

The rest of the paper is organized as follows. Section 2 details the knowledge distillation techniques chosen. In Section 3, the neural architectures designed for black-box VA modeling are presented. Section 4 explains the setup of our experiments to determine the effect of knowledge distillation on the accuracy and size of our distortion models. Section 5 presents the experiment results, while Section 6 concludes the paper.

All the source code for this paper, along with audio examples and trained models, is available online².

2 Knowledge Distillation For Neural Networks

Knowledge distillation is a technique used to compress complex and large "teacher" networks into smaller "student" networks [17, 16]. Knowledge distillation and other model compression techniques offer ways to minimize the computational expenses of working with neural networks, as well as to optimize models for deployment and gain real-time performance advantages.

There exist three primary types of knowledge distillation for neural networks, each with a unique approach to the distillation process: response-based, feature-based, and relation-based distillation. In our study, we use feature-based knowledge distillation, a suitable and popular type of Knowledge Distillation for regression tasks that explores compression of neural audio distortion effects [23, 24].

2.1 Feature-based Distillation

The process of feature-based knowledge distillation is characterized by the transfer of internal representations, or features, of a teacher network to a smaller student network. By retaining these high-level feature extraction capabilities, feature-based distillation allows student models to capture more nuanced representations

¹<https://www.audiority.com/shop/dr-drive/>

²<https://github.com/aleksati/kd-audio-fx>

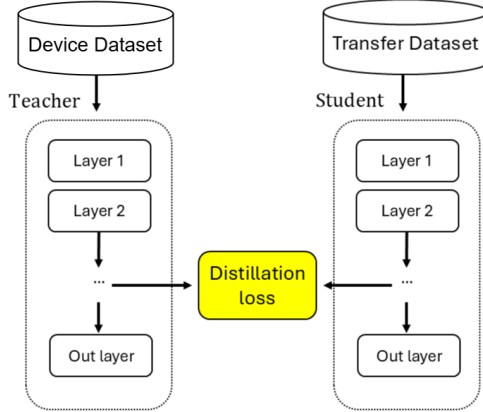


Fig. 1: Feature-based distillation. During training, the feature representations from the last LSTM layers of both the teacher and student models are aligned using a designated distillation loss function.

of the input data, leading to improved generalization and performance.

With feature-based distillation, specific layers of the teacher model are identified as feature extractors that capture important intermediary feature representations of the input data. The intermediate layers of a student network are then aligned to those of the teacher. During training, the student model learns to produce feature maps similar to the teacher model, inheriting its structure and hierarchical understanding of the data. The loss function is defined to minimize the difference between the feature representations of the teacher and student models at these corresponding layers, as illustrated in Figure 1. Traditionally, this process employs the Kullback–Leibler (KL) divergence to measure the difference between the probability distributions derived from the feature representation vectors [16]. However, alternative loss functions can also be utilized [25]. Additionally, this approach can be combined with additional learning to match the output predictions.

3 Neural Network Architecture

Our proposed distillation architecture utilizes recurrent LSTM neural networks, a well-researched and effective

framework for building neural models of distortion and other audio effects [7, 11, 12].

Further, we consider two architectural variations: the depth of the networks and their size. The first teacher model, denoted as H , consists of 7 LSTM layers configured in an hourglass shape, with 8, 16, 32, 64, 32, 16, and 8 units, as illustrated in Figure 2. This design was determined through preliminary experiments, aiming to establish a versatile teacher model that consistently outperforms typically trained student models across all datasets. A second teacher model, denoted as S , consists of 2 LSTM layers, with 64 and 8 units, respectively.

The student models consist of 2 LSTM layers, where the first layer can be configured to be 8, 16, 32, or 64 units, and the final layer is fixed at 8 units to match the final layer of the teacher model. These models will be referred to $S-64-8$, $S-32-8$, $S-16-8$, $S-8-8$. An additional student model, referred to as $S-8$, is configured with only a single 8-unit LSTM layer.

In all models, the input and output layers consist of a linear fully connected layer with a single unit, each processing 1 sample of audio per iteration during inference. Additionally, a skip connection is employed, linking the input directly to the output layer. This design helps integrate raw input features into the final predictions, providing an alternative path for the gradient and improving model convergence.

The conditioning layer is designed to influence the model by incorporating the user-controllable ‘Drive’ knob parameter, which introduces a gradual increase in distortion, as found in the DrDrive plugin. The conditioning process utilizes Feature-wise Linear Modulation (FiLM) and the Gated Linear Unit (GLU) [26, 22]. This layer employs conditioning information to calculate coefficients that modulate the internal representation of the input signal. The output is then passed to the GLU, functioning as a logical gate. The conditioning layer is positioned before the final 8-unit LSTM layer. Consequently, the alignment of feature representations occurs after the conditioning information has already modulated the signal.

4 Experiment Design

In this section, we provide a detailed overview of the dataset and the training strategy employed.

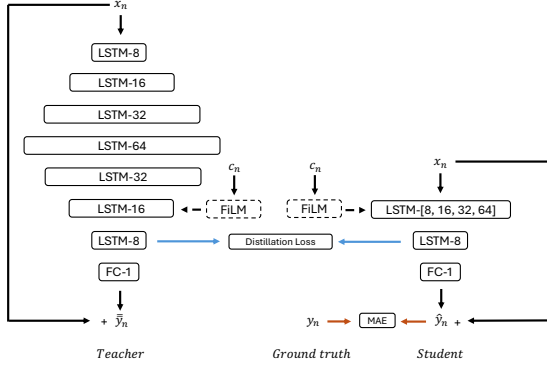


Fig. 2: Teacher H in the hourglass shape (left) and student (right) architectures. The teacher model features 7 LSTM layers, while the student model has only 2 LSTM layers. A conditioning layer is introduced for the parametric case, and it is positioned before the final LSTM layer in both the teacher and student models.

4.1 Dataset

The modeled devices include the Blackstar HT-1 vacuum tube amplifier (HT-1), Electro-Harmonix Big Muff (Big Muff) guitar pedal, and the analog-modeled overdrive plugin DrDrive. The HT-1 is a single-watt dual-channel 8" tube amplifier with classic valve harmonics that offers natural overdrives through gain controls and its "Infinite Shape Feature" (ISF). The Big Muff is a famous distortion guitar pedal effect that produces heavily distorted sounds through its dual cascaded diode clipping stages. The pedal features sustain and volume (pre- and post-gain) control, tone adjustment, and an on/off switch. Finally, DrDrive is a free VST plugin from Audiority³ that offers an analog simulation of an overdrive pedal. Not counting the extended functionalities, DrDrive features four parameter controls for drive, brightness, attack, and level (gain).

Our distillation architecture was evaluated on four datasets, one for each of the HT-1 and Big Muff devices, and two for DrDrive. The HT-1 and Big Muff datasets were identical to those in [15, 9], with the same technical and training configurations. For DrDrive, both parametric and non-parametric sets were considered.

The DrDrive datasets consist of mono audio files with a 10-minute duration, sampled at 48 KHz, where the ef-

³<https://www.audiority.com/shop/dr-drive/>

fect was added to a variety of musical sources ranging from frequency sweeps and white noise with various amplitudes, to physical recordings of musical instruments, such as bass, drums, piano, pads and sections from various pop/rock songs. In preparation for training, 10% of the datasets (1 min) were allocated for the test and validation sets each. In the case of the parametric dataset, the full range of the four VST control parameters was first scaled to the range $[0, 1]$. Then, three different values of the Drive amount were considered as the parametric conditioning (0, 0.5, and 1). For each conditioning instance, fixed values were assigned to the three other VST parameters, attack (0), level (0.3), and brightness (0.5). The datasets were collected using dedicated software for generating datasets for audio-related ML tasks, namely the DGMD presented in [27].

4.2 Training Strategy

Training is conducted for up to 1,000 epochs, with an early stopping mechanism in place to halt training if there is no reduction in validation loss for 30 consecutive epochs. We utilize the Adam optimizer, starting with an initial learning rate of 3×10^{-4} . This learning rate is halved if no improvements are seen after 5 epochs. The batch size is set to 8 examples, with each example composed of 2,048 input audio samples. These configurations are used within the framework of Truncated Backpropagation Through Time (TBPTT). The same strategy is applied in training both the student and teacher models. Test losses and evaluation metrics are computed using the model's weights, which minimize the validation loss throughout the training epochs.

4.3 Loss Function

The teacher and student models included in the experiments are trained using the Mean Absolute Error (MAE) as a loss function. The distillation employed a combination of two loss functions. The feature representations from the last LSTM layers are aligned using the KL divergence of the probability distribution retrieved from the feature representation vectors:

$$\frac{e^{z_i}}{\sum_j e^{z_j}} \quad (1)$$

where z_i represents a single element in the feature vector, and the denominator is the sum of the exponentials

of all elements in the feature vector. This loss function is denoted as L_D . The feature representations from the last teacher’s LSTM layers are pre-computed to be used as targets during the learning process.

The second loss function L_{MAE} is based on MAE, which is computed from the model’s output and the ground truth. These two loss functions are combined into a single loss value, with each component weighted by factors that change over time:

$$L = \alpha(e) \cdot L_D + \beta(e) \cdot L_{MAE} \quad (2)$$

where e represents the epoch number. Specifically, α initial values is 1. and β is 0. α decrease of 10% each epoch, while β is computed as $1 - \alpha$. This weighting strategy enables dynamic optimization, balancing the emphasis primarily on accurately mimicking the teacher’s feature representations and subsequently achieving precise output predictions.

4.4 Evaluation Metrics

To assess the accuracy of the model, we also computed the error-to-signal-ratio (ESR):

$$\frac{\sum_{n=0}^{N-1} |y[n] - \hat{y}[n]|^2}{\sum_{n=0}^{N-1} |y[n]|^2} \quad (3)$$

where y and $\hat{y}$ are the target and prediction, respectively, and the multi-resolution STFT error, which is defined as:

$$\begin{aligned} & \frac{1}{N} \sum_m \left\| \frac{|STFT_m(y)| - |STFT_m(\hat{y})|}{|STFT_m(y)|} \right\|_1 \\ & + \frac{1}{N} \sum_m \left\| \log(|STFT_m(y)|) - \log(|STFT_m(\hat{y})|) \right\|_1 \end{aligned} \quad (4)$$

where $m = [256, 512, 1024]$.

5 Results

Table 1 presents the test loss results of the hourglass-type H teacher models, while Table 2 displays the performance of both distilled and non-distilled student models, S-64-8, S-32-8, S-16-8, S-8-8. Overall, these results highlight the positive impact of distillation, with teacher models achieving the highest accuracies.

We observe that student models typically exhibit improved performance with larger hidden layer sizes,

meaning that the effectiveness of distillation is proportional to the size of the student models, as can be expected. Only the parametric DrDrive case, which incorporates the user-controllable ‘Drive’ knob as a parameter to condition the audio processing, shows a different trend, performing better with a lower size. The parametric results show that distillation works when aligning the feature after conditioning, a task that introduces additional complexity. Also, the non-parametric set of DrDrive and Big Muff exhibit negative improvements at 8 units, but quickly recover at 16 and beyond, following an upward trend.

The HT-1 dataset emerges as the most challenging case, as it produces the largest errors when compared to other datasets. On the other hand, HT-1 demonstrates the most significant improvements through distillation. In contrast, when non-distilled students are already performing well, the benefits of distillation are less pronounced.

Table 2 also reports the number of parameters and related floating-point operations per sample (FLOP). Compared to teacher models, the FLOP count shows that student models require significantly fewer parameters and operations per sample, representing a big improvement in memory requirements and efficiency.

Figure 3 takes a closer look at the performance advantages of distilled versus non-distilled models with respect to the size of their hidden layers. STFT and MAE loss show relatively stable performance across all models, following an upward trend. In contrast, ESR results exhibit greater variability, with both extreme negative and positive values. This greater variance could be because ESR is a normalized function, and the distillation improves more when the signal’s amplitude is larger than low. Lastly, HT-1 and parametric DrDrive datasets show larger improvements at lower sizes, an opposite trend compared to the other datasets.

The results further pinpoint that the H teacher model is not always effective for the smallest students. Table 3 presents the results of our experiments using the teacher S model, which exhibits a smaller complexity gap compared to the students. This scenario shows improvements even in the smallest student models. In particular, both DrDrive and Big Muff demonstrated advantages from distillation, as well as HT-1, although to a slightly lesser extent. Similar to the H teacher scenario, the performance improvements of students

Table 1: Test performance and loss of H teacher models with related number of parameters and floating-point operations per sample (FLOPs).

Device dataset	Total params	FLOPs	MSE	MAE	ESR	STFT
HT-1	49385	98768	$4.20 \cdot 10^{-3}$	$3.92 \cdot 10^{-2}$	4.11	0.15
Big Muff	-	-	$4.00 \cdot 10^{-4}$	$5.30 \cdot 10^{-3}$	1.48	0.24
DrDrive (un-cond)	-	-	$3.00 \cdot 10^{-4}$	$1.08 \cdot 10^{-2}$	0.63	0.10
DrDrive (cond)	49993	98898	$6.71 \cdot 10^{-5}$	$4.63 \cdot 10^{-3}$	0.15	0.05

Table 2: Test performance and loss of student models (S-64-8, S-32-8, S-16-8, S-8-8) trained on H teacher model, with related number of parameters and floating-point operations per sample (FLOPs). Instances with the best performance are highlighted in bold. Variations in the hidden size of the network’s LSTM layer are presented row-by-row.

Device dataset	Student	Total params	FLOP count	MSE non-distilled	distilled	MAE non-distilled	distilled	ESR non-distilled	distilled	STFT non-distilled	distilled
HT-1	S-8-8	873	1744	$4.91 \cdot 10^{-2}$	$4.97 \cdot 10^{-2}$	$1.44 \cdot 10^{-1}$	$1.34 \cdot 10^{-1}$	41.97	24.92	0.56	0.59
	S-16-8	1961	3856	$2.72 \cdot 10^{-2}$	$1.34 \cdot 10^{-2}$	$1.07 \cdot 10^{-1}$	$7.78 \cdot 10^{-2}$	23.72	11.17	0.42	0.30
	S-32-8	5673	11152	$1.17 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	$6.65 \cdot 10^{-2}$	$6.34 \cdot 10^{-2}$	12.51	8.96	0.26	0.24
	S-64-8	19241	38032	$9.47 \cdot 10^{-3}$	$8.69 \cdot 10^{-3}$	$5.99 \cdot 10^{-2}$	$5.85 \cdot 10^{-2}$	8.60	6.73	0.24	0.23
Big Muff	S-8-8	873	1744	$1.03 \cdot 10^{-3}$	$9.75 \cdot 10^{-4}$	$9.88 \cdot 10^{-3}$	$1.38 \cdot 10^{-2}$	5.41	6.00	0.47	0.49
	S-16-8	1961	3856	$1.10 \cdot 10^{-3}$	$8.83 \cdot 10^{-4}$	$1.11 \cdot 10^{-2}$	$1.35 \cdot 10^{-2}$	4.90	5.32	0.60	0.48
	S-32-8	5673	11152	$1.06 \cdot 10^{-3}$	$8.86 \cdot 10^{-4}$	$1.01 \cdot 10^{-2}$	$9.31 \cdot 10^{-3}$	4.92	2.69	0.51	0.47
	S-64-8	19241	38032	$5.50 \cdot 10^{-4}$	$4.39 \cdot 10^{-4}$	$6.47 \cdot 10^{-3}$	$5.70 \cdot 10^{-3}$	1.82	1.24	0.32	0.27
DrDrive (un-cond)	S-8-8	873	1744	$2.97 \cdot 10^{-3}$	$2.59 \cdot 10^{-3}$	$3.38 \cdot 10^{-2}$	$3.08 \cdot 10^{-2}$	5.92	6.19	0.29	0.23
	S-16-8	1961	3856	$1.20 \cdot 10^{-3}$	$1.00 \cdot 10^{-3}$	$1.97 \cdot 10^{-2}$	$1.84 \cdot 10^{-2}$	1.97	1.87	0.17	0.15
	S-32-8	5673	11152	$1.08 \cdot 10^{-3}$	$7.00 \cdot 10^{-4}$	$1.91 \cdot 10^{-2}$	$1.59 \cdot 10^{-2}$	2.00	1.35	0.16	0.13
	S-64-8	19241	38032	$6.81 \cdot 10^{-4}$	$4.53 \cdot 10^{-4}$	$1.49 \cdot 10^{-2}$	$1.23 \cdot 10^{-2}$	1.16	0.90	0.13	0.10
DrDrive (cond)	S-8-8	1049	1810	$3.63 \cdot 10^{-3}$	$3.30 \cdot 10^{-3}$	$3.31 \cdot 10^{-2}$	$3.15 \cdot 10^{-2}$	7.74	3.68	0.30	0.26
	S-16-8	2569	3986	$2.67 \cdot 10^{-3}$	$2.36 \cdot 10^{-3}$	$2.88 \cdot 10^{-2}$	$2.78 \cdot 10^{-2}$	5.13	3.69	0.24	0.22
	S-32-8	7913	11410	$3.70 \cdot 10^{-3}$	$3.57 \cdot 10^{-3}$	$3.27 \cdot 10^{-2}$	$3.21 \cdot 10^{-2}$	7.11	4.22	0.30	0.26
	S-64-8	27817	38546	$4.42 \cdot 10^{-3}$	$3.84 \cdot 10^{-3}$	$3.46 \cdot 10^{-2}$	$3.29 \cdot 10^{-2}$	8.02	4.60	0.30	0.25

Table 3: Test performance and loss of student models (S-32-8, S-16-8, S-8-8) trained on S teacher model, with the related number of parameters and floating-point operations per sample (FLOPs). Instances with the best performance are highlighted in bold. Variations in the hidden size of the network’s LSTM layer are presented row-by-row.

Device dataset	Student	Total params	FLOP count	MSE non-distilled	distilled	MAE non-distilled	distilled	ESR non-distilled	distilled	STFT non-distilled	distilled
HT-1	S-8-8	873	1744	$4.91 \cdot 10^{-2}$	$4.10 \cdot 10^{-2}$	$1.44 \cdot 10^{-1}$	$1.36 \cdot 10^{-1}$	41.97	30.65	0.54	0.49
	S-16-8	1961	3856	$2.72 \cdot 10^{-2}$	$2.48 \cdot 10^{-2}$	$1.07 \cdot 10^{-1}$	$1.04 \cdot 10^{-1}$	23.72	20.10	0.42	0.40
	S-32-8	5673	11152	$1.17 \cdot 10^{-2}$	$1.72 \cdot 10^{-2}$	$6.65 \cdot 10^{-2}$	$8.35 \cdot 10^{-2}$	12.51	13.46	0.26	0.33
Big Muff	S-8-8	873	1744	$1.00 \cdot 10^{-3}$	$1.00 \cdot 10^{-3}$	$1.00 \cdot 10^{-2}$	$1.00 \cdot 10^{-2}$	5.41	3.83	0.47	0.45
	S-16-8	1961	3856	$1.10 \cdot 10^{-3}$	$9.99 \cdot 10^{-4}$	$1.11 \cdot 10^{-2}$	$9.98 \cdot 10^{-3}$	4.90	3.86	0.60	0.45
	S-32-8	5673	11152	$5.50 \cdot 10^{-4}$	$5.19 \cdot 10^{-4}$	$1.01 \cdot 10^{-2}$	$6.48 \cdot 10^{-3}$	4.92	1.85	0.51	0.32
DrDrive (un-cond)	S-8-8	873	1744	$2.40 \cdot 10^{-3}$	$2.23 \cdot 10^{-3}$	$2.98 \cdot 10^{-2}$	$2.81 \cdot 10^{-2}$	5.92	5.12	0.23	0.23
	S-16-8	1961	3856	$1.20 \cdot 10^{-3}$	$5.45 \cdot 10^{-4}$	$1.97 \cdot 10^{-2}$	$1.34 \cdot 10^{-2}$	1.97	0.96	0.17	0.11
	S-32-8	5673	11152	$1.08 \cdot 10^{-3}$	$5.76 \cdot 10^{-4}$	$1.91 \cdot 10^{-2}$	$1.34 \cdot 10^{-2}$	2.00	0.93	0.16	0.11
DrDrive (cond)	S-8-8	1049	1810	$3.63 \cdot 10^{-3}$	$3.12 \cdot 10^{-3}$	$3.31 \cdot 10^{-2}$	$3.09 \cdot 10^{-2}$	7.74	3.50	0.30	0.26
	S-16-8	2569	3986	$2.67 \cdot 10^{-3}$	$3.02 \cdot 10^{-3}$	$2.88 \cdot 10^{-2}$	$3.05 \cdot 10^{-2}$	5.13	3.42	0.24	0.25
	S-32-8	7913	11410	$3.70 \cdot 10^{-3}$	$9.60 \cdot 10^{-4}$	$3.27 \cdot 10^{-2}$	$2.11 \cdot 10^{-2}$	7.11	1.45	0.30	0.27

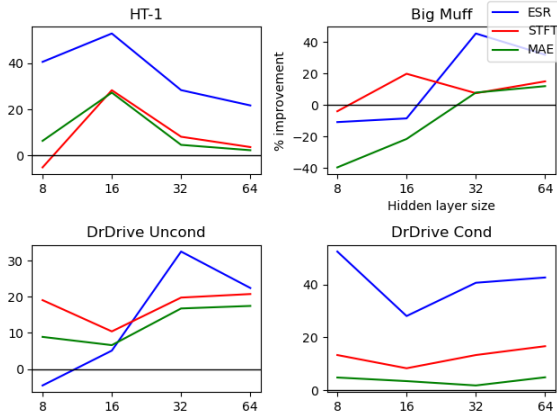


Fig. 3: Plots detailing the improvements, in percent (%), of using distilled models over non-distilled models for different hidden layer sizes, with the H teacher. Comparisons are made using ESR, STFT, and MAE loss metrics.

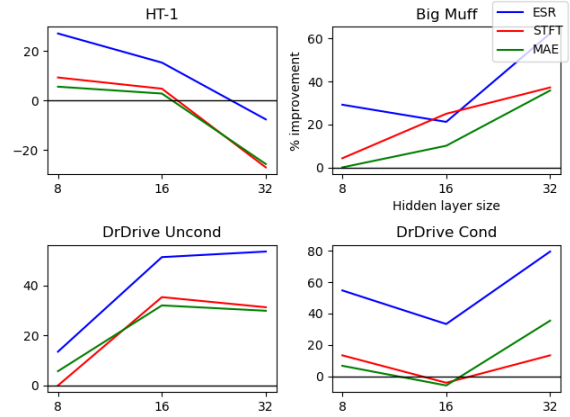


Fig. 4: Plots detailing the improvements, in percent (%), of using distilled models over non-distilled models for different hidden layer sizes, with the S teacher. Comparisons are made using ESR, STFT, and MAE loss metrics.

distilled from a teacher S model tend to be more significant as their size increases. This trend is illustrated in Figure 4, which compares the performance of distilled and non-distilled models in relation to the size of their hidden layers. The improvements in STFT and MAE metrics are nearly identical, while the ESR results closely mirror these trends. Notably, except for the HT-1 dataset, the performance gains observed with distilled models are greater than those achieved in the H teacher case. Moreover, these improvements tend to increase as the model sizes grow.

Finally, Table 4 presents the results for a distilled student model with a single LSTM layer ($S-8$) trained on both the H and S teachers. These cases demonstrate that a smaller complexity gap between teachers and students typically results in enhanced performance, showing that distillation is effective in the majority of cases. However, the HT-1 dataset does not show significant improvements. We also observe that the parametric DrDrive student performs best when trained by the H teacher model.

5.1 Listening tests

A MUSHRA-style listening test was conducted to investigate the perceptual quality of distilled versus non-distilled models [28]. Audio examples from all three

device datasets were included, along with parametric examples of DrDrive. However, only audio from $S-64-8$ student models trained on the H teacher model was considered, as these models achieved the lowest test loss on average. In addition, we included audio examples from $S-8-8$ student models trained on the S teacher model.

The listening test included two sets of audio examples per device model, comprising both distilled and non-distilled examples. When considering both the $S-64-8$ and $S-8-8$ cases, this amounted to sixteen conditions total for one listening trial. In each trial condition, five audio examples and one reference were presented to participants, consisting of the audio distortion from the device effects applied to either guitar, bass, or drum signals. Participants were asked to rate the perceptual quality of each audio example against the reference on a scale from 0 to 100. The reference audio acted as the ground truth and was therefore sourced directly from the device dataset. The remaining five audio examples included a hidden reference, low and mid-range anchors, where reference duplicates are passed through low-pass filters at 3.5KHz and 7KHz, respectively, as well as the outputs of the distilled and non-distilled models.

Of the thirteen people who participated in the listening tests, three were removed in the post-screening as they

Table 4: Test performance and loss of student models (S-8) trained on H and S teacher models, with the related number of parameters and floating-point operations per sample (FLOPs) are reported. Instances with the best performance are highlighted in bold.

Device dataset	Teacher	Total params	FLOP count	MSE		MAE		ESR		STFT	
				non-distilled	distilled	non-distilled	distilled	non-distilled	distilled	non-distilled	distilled
HT-1	S	329	656	$5.03 \cdot 10^{-2}$	$5.47 \cdot 10^{-2}$	$1.50 \cdot 10^{-1}$	$1.57 \cdot 10^{-1}$	39.19	30.07	0.61	0.61
-	H	-	-	-	$5.58 \cdot 10^{-2}$	-	$1.60 \cdot 10^{-1}$	-	30.99	-	0.61
Big Muff	S	-	-	$1.00 \cdot 10^{-3}$	$9.87 \cdot 10^{-4}$	$9.83 \cdot 10^{-3}$	$9.83 \cdot 10^{-3}$	5.57	4.12	0.51	0.48
-	H	-	-	-	$1.00 \cdot 10^{-3}$	-	$9.97 \cdot 10^{-2}$	-	4.02	-	0.51
DrDrive (un-cond)	S	-	-	$3.98 \cdot 10^{-3}$	$1.28 \cdot 10^{-3}$	$3.97 \cdot 10^{-2}$	$1.83 \cdot 10^{-2}$	6.97	1.07	0.33	0.20
-	H	-	-	-	$3.07 \cdot 10^{-3}$	-	$3.45 \cdot 10^{-1}$	-	7.51	-	0.26
DrDrive (cond)	S	337	666	$1.60 \cdot 10^{-3}$	$1.26 \cdot 10^{-3}$	$3.60 \cdot 10^{-2}$	$2.41 \cdot 10^{-2}$	3.75	3.41	0.32	0.28
-	H	-	-	-	$1.23 \cdot 10^{-3}$	-	$2.39 \cdot 10^{-2}$	-	3.25	-	0.27

scored the hidden reference below 80 in more than 15% of trial conditions. The trials were conducted using a pair of Beyerdynamic DT 770 Pro headphones.

As presented in Figure 5, results show high overall scores for distilled and non-distilled examples. However, the perceptual improvements of distillation are not pronounced, as non-distilled models already start with high accuracy. This finding was further substantiated by p -values gathered from statistical T-tests comparing listening test scores from distilled and non-distilled models. On all devices, the p -value scores were greater than 0.05, meaning it’s still inconclusive whether there are definitive perceptual advantages of distilled models over non-distilled models. On the other hand, we observe a slight preference for distilled models in most cases, especially in the S-64-8 students. Non-distilled models also show greater variability, suggesting slightly less consistency in their perceived quality over their distilled counterparts.

6 Conclusion

The paper explores knowledge distillation techniques to reduce the size of large neural network models and enhance the accuracy of smaller ones in the context of emulating audio distortion effects. In particular, the study examines feature-based knowledge distillation for audio-to-audio modeling on both non-parametric and parametric datasets, using an LSTM architecture to train smaller student models that replicate the internal representations of larger teacher networks. Our results indicate that the process of distillation can be used to optimize training for VA audio models, as we notice improved model accuracy when comparing distilled and non-distilled networks of equal parameter size.

Notably, the distillation improvements are more pronounced as student models increase in size. The results also highlight the known challenges of using feature-based distillation techniques, demonstrating that it is more difficult for models to learn from teachers that present particularly more complexity. Finally, listening tests provided additional qualitative data on the perceptual effects of distillation for audio effects modeling. Although not statistically conclusive, the listening test results show positive tendencies toward using distillation to improve the perceptual quality of audio effect models.

For future work, more exhaustive architectural modifications should be considered. Additionally, designing more efficient loss functions could help refine the distance metrics between the probability distributions of teachers and students. Finally, employing a multi-teacher ensemble approach could be advantageous in mitigating potential complexity gaps between models.

References

- [1] Pekonen, J. and Välimäki, V., “The brief history of virtual analog synthesis,” in *Proc. 6th Forum Acusticum. Aalborg, Denmark: European Acoustics Association*, 2011.
- [2] Rest, M., Parker, J. D., and Werner, K. J., “WDF Modeling of a Korg MS-50 Based Non-linear Diode Bridge VCF,” in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2017.
- [3] Werner, K. J., Bernardini, A., Smith, J. O., and Sarti, A., “Modeling Circuits With Arbitrary

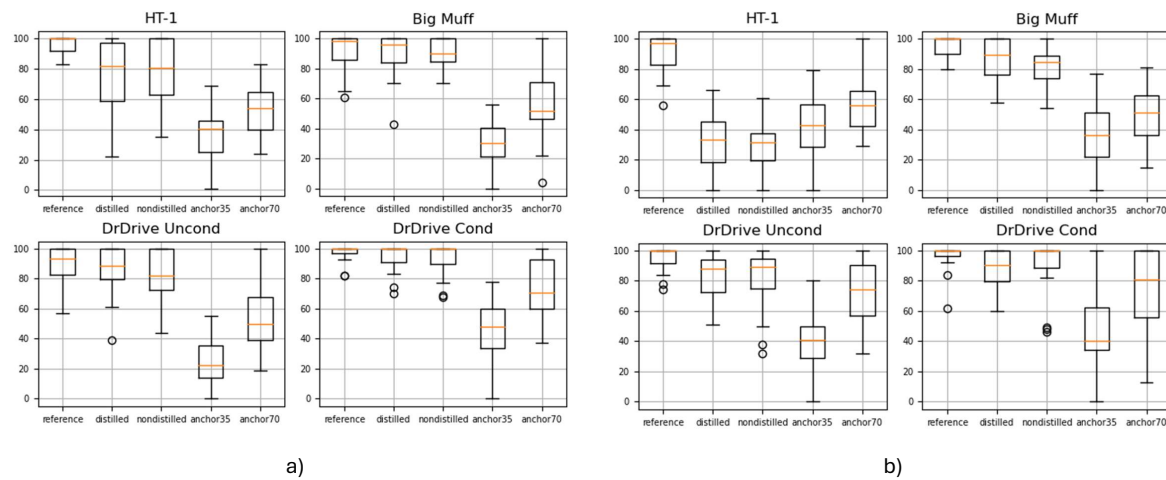


Fig. 5: Box plot showing the rating results of the listening tests for S-64-8 students trained on H teacher (a) and S-8-8 students trained on S teacher (b), for HT-1 (top-left), Big Muff (top-right), non-parametric DrDrive (bottom-left), and parametric DrDrive (bottom-right). The orange horizontal lines indicate the median value, and the outliers are displayed as circles.

- Topologies and Active Linear Multiports Using Wave Digital Filters,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, 65(12), 2018, doi:10.1109/TCSI.2018.2837912.
- [4] Eichas, F. and Zölzer, U., “Black-box modeling of distortion circuits with block-oriented models,” in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2016.
- [5] Holmes, B. and Van Walstijn, M., “Physical model parameter optimisation for calibrated emulation of the Dallas Rangemaster Treble Booster guitar pedal,” in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2016.
- [6] Damskägg, E.-P., Juvela, L., Thuillier, E., and Välimäki, V., “Deep learning for tube amplifier emulation,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.
- [7] Covert, J. and Livingston, D., “A vacuum-tube guitar amplifier model using a recurrent neural network,” in *Proceedings of IEEE Southeastcon*, 2013.
- [8] Damskägg, E.-P., Juvela, L., and Valimäki, V., “Real-Time Modeling of Audio Distortion Circuits with Deep Learning,” in *Proceedings of the Sound and Music Computing conference (SMC)*, 2019.
- [9] Wright, A., Damskägg, E.-P., and Välimäki, V., “Real-Time Black-Box Modelling With Recurrent Neural Networks,” in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2019.
- [10] Steinmetz, C. J. and Reiss, J. D., “Efficient neural networks for real-time modeling of analog dynamic range compression,” in *152nd Convention of the Audio Engineering Society (AES)*, 2022.
- [11] Simionato, R. and Fasciani, S., “Modeling Time-Variant Responses of Optical Compressors with Selective State Space Models,” *Journal of Audio Engineering Society*, 73(3), pp. 144–165, 2025.
- [12] Wright, A. and Välimäki, V., “Neural Modeling of Phaser and Flanging Effects,” *Journal of the Audio Engineering Society*, 69(7/8), 2021, doi: 10.17743/jaes.2021.0029.
- [13] Mikkonen, O., Wright, A., Moliner, E., and Välimäki, V., “Neural modeling of magnetic tape

- recorders,” *International Conference on Digital Audio Effects (DAFx)*, 2023.
- [14] Marinó, G. C., Petrini, A., Malchiodi, D., and Frasca, M., “Deep neural networks compression: A comparative survey and choice recommendations,” *Neurocomputing*, 520, 2023.
- [15] Südholt, D., Wright, A., Erkut, C., and Välimäki, V., “Pruning Deep Neural Network Models of Guitar Distortion Effects,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31, 2023, doi:10.1109/TASLP.2022.3223257.
- [16] Hinton, G., Vinyals, O., and Dean, J., “Distilling the Knowledge in a Neural Network,” in *NIPS Deep Learning and Representation Learning Workshop*, 2015.
- [17] Gou, J., Yu, B., Maybank, S. J., and Tao, D., “Knowledge Distillation: A Survey,” *International Journal of Computer Vision*, 129(6), 2021, doi:10.1007/s11263-021-01453-z.
- [18] Kang, Z., He, Y., Wang, J., Peng, J., Qu, X., and Xiao, J., “Feature-Rich Audio Model Inversion for Data-Free Knowledge Distillation Towards General Sound Classification,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.
- [19] Choi, K., Kersner, M., Morton, J., and Chang, B., “Temporal Knowledge Distillation for on-device Audio Classification,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.
- [20] Gao, L., Xu, K., Wang, H., and Peng, Y., “Multi-representation knowledge distillation for audio classification,” *Multimedia Tools and Applications*, 81(4), 2022, doi:10.1007/s11042-021-11610-8.
- [21] Gholami, B., El-Khamy, M., and Song, K., “Knowledge Distillation for Tiny Speech Enhancement with Latent Feature Augmentation,” in *Proceedings of Interspeech*, 2024.
- [22] Simionato, R. and Fasciani, S., “Conditioning Methods for Neural Audio Effects,” in *Proceedings of the Sound and Music Computing conference (SMC)*, 2024.
- [23] Passalis, N. and Tefas, A., “Learning deep representations with probabilistic knowledge transfer,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.
- [24] Xu, Q., Wu, K., Wu, M., Mao, K., Li, X., and Chen, Z., “Reinforced Knowledge Distillation for Time Series Regression,” *IEEE Transactions on Artificial Intelligence*, 2023.
- [25] Park, W., Kim, D., Lu, Y., and Cho, M., “Relational Knowledge Distillation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [26] Perez, E., Strub, F., De Vries, H., Dumoulin, V., and Courville, A., “Film: Visual reasoning with a general conditioning layer,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.
- [27] Fasciani, S., Simionato, R., and Tidemann, A., “A Universal Tool for Generating Datasets from Audio Effects,” in *Proceedings of the Sound and Music Computing conference (SMC)*, 2024.
- [28] Schoeffler, M., Bartoschek, S., Stöter, F.-R., Roess, M., Westphal, S., Edler, B., and Herre, J., “webMUSHRA—A comprehensive framework for web-based listening tests,” *Journal of Open Research Software*, 6(1), 2018.