НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
март–апрель 2026 Том 26 № 2 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
March–April 2026 Vol. 26 No 2 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2026-26-2-295-305

Spectral-based multi-band recurrent neural networks for black-box modeling of dynamic range compressors

Andrei F. Balykin[1]✉, Ivan S. Blekanov[2]

[1,2] St. Petersburg State University (SPbU), Saint Petersburg, 199034, Russian Federation

[1] st054659@student.spbu.ru✉, https://orcid.org/0009-0003-1554-2873
[2] i.blekanov@spbu.ru, https://orcid.org/0000-0002-7305-1429

Abstract

Deep learning approaches have been increasingly adopted for virtual analog modeling, which aims to replicate the sonic characteristics of analog audio devices. In the context of analog dynamic range compressor modeling, many existing methods operate directly on raw audio waveforms which are high-dimensional and contain fine-grained temporal features at high sampling rates. These representations are computationally demanding and limit model efficiency. We propose a feature extraction pipeline that leverages the magnitude component of the Short-Time Fourier Transform in combination with a spectral amplification mechanism which acts similarly to a spectral mask but can both attenuate and amplify selected frequency components. We employ multi-band Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures that split the magnitude spectrum into several frequency bands for independent processing, substantially reducing computational complexity while preserving high modeling accuracy. To evaluate our approach, we created two datasets consisting of recordings of the consumer-grade analog compressor Alesis 3630 and its digital counterpart, discoDSP NightShine. We conducted extensive experiments comparing our method against raw waveform baselines using four objective metrics, theoretical and empirical measurements of computational performance, and a subjective listening test. Results indicate that single-band models based on the proposed feature extraction pipeline outperform raw-audio baselines across all evaluation metrics. Multi-band configurations further improve the efficiency balance. In particular, four-band LSTM and GRU architectures achieve higher perceptual fidelity at substantially lower computational cost. Moreover, we conducted a subjective listening test that yielded results aligned with the objective metrics. All source code and pretrained models are provided for reproducibility.

Keywords

signal processing, deep learning, virtual analog modeling, black-box modeling, recurrent neural networks

For citation: Balykin A.F., Blekanov I.S. Spectral-based multi-band recurrent neural networks for black-box modeling of dynamic range compressors. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2026, vol. 26, no. 2, pp. 295–305. doi: 10.17586/2226-1494-2026-26-2-295-305

УДК 004.032.26

Спектральные многополосные рекуррентные нейронные сети для моделирования компрессоров динамического диапазона методом «черного ящика»

Андрей Федорович Балыкин[1]✉, Иван Станиславович Блеканов[2]

[1,2] Санкт-Петербургский государственный университет, Санкт-Петербург, 199034, Российская Федерация

[1] st054659@student.spbu.ru✉, https://orcid.org/0009-0003-1554-2873
[2] i.blekanov@spbu.ru, https://orcid.org/0000-0002-7305-1429

Аннотация

Введение. Подходы глубокого обучения все активнее применяются для задач виртуального аналогового моделирования, цель которых заключается в воспроизведении звуковых характеристик аналоговых аудиоустройств. В области моделирования аналоговых компрессоров динамического диапазона многие существующие методы работают с аудиосигналами во временной области, что обуславливает высокую

размерность входного сигнала при высокой частоте дискретизации. Обработка таких высокодетализированных признаков является вычислительно затратной и снижает эффективность моделей. **Метод.** Представлен метод предварительной обработки признаков, использующий амплитудную компоненту кратковременного преобразования Фурье в сочетании с механизмом спектрального усиления, функционирующим аналогично спектральной маске, но способным как ослаблять, так и усиливать частотные компоненты. В качестве рассматриваемых архитектур были предложены многополосные сети Long Short-Term Memory (LSTM) и Gated Recurrent Unit (GRU), которые разделяют амплитудный спектр на несколько частотных полос для независимой обработки, что существенно снижает вычислительную сложность при сохранении высокой точности моделирования. **Основные результаты.** Для оценки представленного подхода были сформированы два набора данных, содержащих записи с аналогового компрессора Alesis 3630 и его цифровой эмуляции discoDSP NightShine. На выбранных наборах данных были проведены эксперименты, в которых предложенный метод сравнивался с базовыми моделями по четырем объективным метрикам, теоретическим и эмпирическим показателям вычислительной производительности, а также результатам субъективного прослушивания. **Обсуждение.** Результаты показали, что однополосные модели с использованием разработанного метода извлечения признаков превосходят базовые модели по всем оценочным метрикам. Многополосные конфигурации обеспечивают более выгодный баланс между качеством и вычислительной эффективностью. В частности, четырехполосные архитектуры LSTM и GRU демонстрируют более высокую перцептивную точность при существенно меньших вычислительных затратах. Кроме того, был проведен субъективный тест прослушивания, результаты которого согласуются с объективными метриками. Исходный код и предобученные модели опубликованы в открытом доступе для обеспечения воспроизводимости результатов.

Ключевые слова
обработка сигналов, глубокое обучение, виртуальное аналоговое моделирование, метод черного ящика, рекуррентные нейронные сети

Ссылка для цитирования: Балыкин А.Ф., Блеканов И.С. Спектральные многополосные рекуррентные нейронные сети для моделирования компрессоров динамического диапазона методом «черного ящика» // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 2. С. 295–305 (на англ. яз.). doi: 10.17586/2226-1494-2026-26-2-295-305

Introduction

Analog audio devices have established the technological foundation for the modern recording industry and remain fundamental tools for music producers and audio engineers due to their high-fidelity processing and distinctive tonal characteristics [1]. However, physical implementations of such devices present significant limitations, including maintenance requirements and high manufacturing costs. In addition, their operation requires dedicated space and complex connectivity when multiple devices are used.

Among these devices, analog compressors remain one of the most widely employed tools in modern studio environments. Compressors are specifically designed to control the dynamic range of audio signals by reducing the volume difference between their loudest and quietest parts. Dynamic range reduction in analog compressors is combined with nonlinear amplitude and phase modifications introduced by internal circuitry, resulting in analog coloration [2].

With advancements in recording technology, analog hardware has been progressively replaced by digital solutions. The integration of Digital Audio Workstations (DAWs) into mainstream audio production has fundamentally reshaped both engineering practices and creative workflows [3]. Central to DAW functionality are digital signal processing algorithms which include digital audio compressors frequently implemented as Virtual Studio Technology (VST) plugins [4]. Compared to analog compression, digital compressors often sound overly clean or sterile due to minimal distortion and limited amplitude and phase modification during dynamic range processing.

Increasing interest in replicating the sound and behavior of analog hardware in digital form has established virtual analog modeling as a major research direction [5]. Traditionally, achieving accurate models of analog devices has required substantial manual effort and expertise. Recent advancements in deep learning, however, have made this process more efficient by enabling neural networks to learn directly from audio recordings of analog equipment [6]. Despite these advancements, many neural network methods work at the raw audio sample level, resulting in challenges, such as limited access to global signal context, reduced generalization across different device parameters, and high computational costs that restrict their practical use.

In response to the computational and dimensional complexity of raw audio data, preprocessing techniques, such as Mel-spectrograms, Mel-Frequency Cepstral Coefficients (MFCCs) [7], and Short-Time Fourier Transform (STFT)-based methods, have been proposed to provide more compact audio features. Furthermore, various approaches have been developed for processing STFT spectral bins, such as using their raw complex values [8], separating magnitude and phase components [9], or relying exclusively on magnitude information with subsequent phase reconstruction through vocoders.

In this paper, we introduce an STFT-based model that leverages magnitude response amplification, a promising approach within virtual analog modeling. Furthermore, we propose a multi-band Recurrent Neural Network (RNN) architecture which partitions the spectral bins into multiple parallel frequency bands for independent processing, enabling an optimal balance between modeling precision and computational complexity. Our study relies on two newly created datasets: one representing the consumer-grade analog compressor Alesis 3630, and another produced using discoDSP NightShine, a digital model of the same device.

Related Work

Research in digital audio effects has shifted toward digital emulation of analog devices, driven by the standardization of computer-based editing software [10]. Digital compressors exemplify the significant influence of algorithmic design on the perceptual qualities of processed audio, often deriving inspiration from analog hardware to replicate the sonic characteristics of classic units [11].

The digital emulation of analog devices, known as virtual analog modeling, relies on two primary strategies defined by the level of available internal knowledge. With full circuit schematics, a white-box methodology can precisely simulate component interactions using techniques like state-space models and wave digital filters [12, 13]. With unknown schematics, the modeling task shifts to a black-box paradigm. The black-box methods address cases in which the internal structure of the analog device is unknown or inaccessible, relying solely on observable inputs, control parameters, and outputs to reproduce system behavior. Common black-box techniques include Wiener models [14] and Volterra series [15] which are widely used to capture input–output relationships.

Recently, deep learning methods have gained popularity in audio processing, showing strong performance in tasks, such as Automatic Speech Recognition, Text-to-Speech synthesis, and Automatic Speaker Verification. The WaveNet architecture [16] introduced causal dilated convolutions which have since been widely adopted in audio applications, including virtual analog modeling. RNNs, particularly LSTM and GRU architectures, have also been used to model analog devices directly from raw waveforms [17]. In addition, fully connected, recurrent, and convolutional neural networks have demonstrated the ability to model various analog hardware systems in real-time [18, 19].

One of the earliest studies on modeling analog compressors with deep neural networks was introduced by SignalTrain [20] which proposed a two-branch architecture that processes magnitude and phase responses separately. Additionally, the authors introduced a dataset of recordings from the LA-2A analog compressor, enabling further research in the field. Following this work, a lightweight Temporal Convolutional Network (TCN) architecture was proposed for modeling the LA-2A compressor utilizing the same SignalTrain LA-2A dataset [21]. Subsequently, an encoder-decoder model based on LSTM networks was introduced in [22], effectively modeling LA-2A and CL-1B analog compressors.

Subsequent research concentrated on optimizing the accuracy and efficiency with the state-space approach. The Structured State-Space (S4) model, employed in [23], offered a high-performance alternative to recurrent architectures, demonstrating superior results compared to the models used in [21]. This was further advanced in [24] with the Selective State-Space (S6) model which provided a more accurate and efficient solution for modeling analog compressors.

Proposed methods

STFT Features

In this work, we leverage STFT transformations to extract frequency-domain features which are used as inputs to neural networks, replacing the direct processing of raw waveforms. Stable reconstruction is provided by the Nonzero Overlap-Add (NOLA) condition, requiring squared and shifted analysis windows to overlap without gaps. In our setup, we used a Hann window with 50 % overlap which satisfies the NOLA condition, enabling inverse STFT reconstruction via least-squares Weighted Overlap-Add (WOLA), where the overlap-added frames are normalized by the window energy. For a discrete signal $x[n]$ the STFT is defined as:

$$X(k, n) = \sum_{m=0}^{N-1} x[nR + m]\, w[m] e^{-j\frac{2\pi}{N}km}, \; k = 0, \ldots, N-1,$$

where, k, n, R, N, m, $w[m]$ are frequency bin index, time frame index, hop size, FFT size, sample index within a frame, and analysis window, respectively.

In audio compression, the core processing operation is a parameterized volume adjustment which mainly affects the magnitude component of the STFT. For simplicity and efficiency, we utilize only the magnitude components of the STFT as input features for the neural network. The original phase is combined with the processed magnitude to reconstruct the final audio signal via inverse STFT (iSTFT). Let the processed magnitude be $\hat{A}(k, n)$, the original phase be $\varphi(k, n) = \arg\{X(k, n)\}$, and the reconstructed complex spectrogram be $\hat{X}(k, n) = \hat{A}(k, n)e^{j\varphi(k,n)}$. The time-domain reconstruction is obtained by least-squares WOLA. We first perform the per-frame inverse Discrete Fourier Transform (DFT), and $y_n[m]$ denotes the reconstructed time-domain frame for the n-th STFT frame:

$$y_n[m] = \frac{1}{N} \sum_{k=0}^{N-1} \hat{X}(k, n)\, e^{+j\frac{2\pi}{N}km}, \; m = 0, \ldots, N-1,$$

which transforms each spectral frame back into the time domain. Then we apply overlap-add with window-sum-of-squares normalization, with ε as a small constant added for numerical stability (set to 10^{-8} in our implementation), resulting in the reconstructed time-domain signal $\hat{x}[m]$:

$$\hat{x}[m] = \frac{\sum_n y_n[m - nR] w[m - nR]}{\sum_n w^2[m - nR] + \varepsilon}.$$

Moreover, we modified the commonly used spectral masking approach employed in music source separation tasks to manipulate the spectral content [8]. Typically, spectral masks are computed within the [0, 1] range, facilitating attenuation of frequency-domain magnitude components derived from the STFT. However, due to the nonlinearities in audio compression devices, our approach requires both attenuation and enhancement of frequency-domain features. Therefore, we introduce a learnable magnitude amplification parameter, allowing the network to dynamically adjust spectral content beyond attenuation. We refer to this approach as spectral amplification and will use this term throughout the rest of the paper. The corresponding pipeline is shown in Fig. 1.

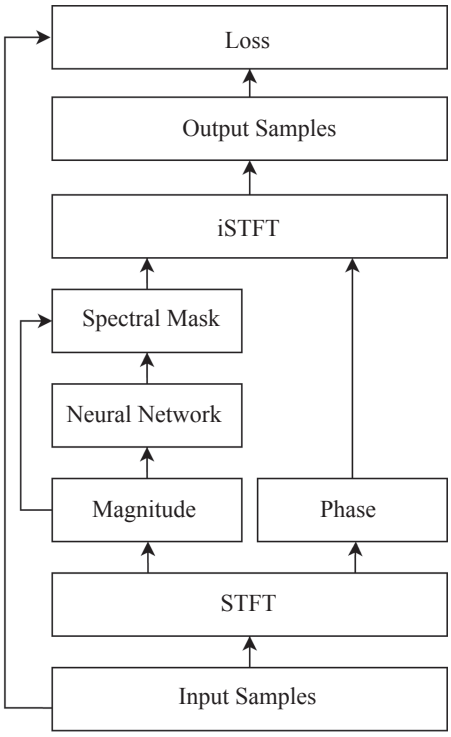

Fig. 1. Schema of the proposed audio processing pipeline

Single-band Model

Both LSTM and GRU networks employ the same input feature shape and can be used interchangeably. In this section, we refer to LSTM and GRU networks as RNNs, meaning that either of these or other similar recurrent architectures could be used in their place. We consider an input sequence of magnitude components obtained from the STFT transform and concatenated with the device parameters vector as an input for a multi-layer RNN (Fig. 2). Then the Multi-Layer Perceptron (MLP) is applied to post-process RNN outputs concatenated with the parameters vector once again to condition the MLP layer on device parameters.

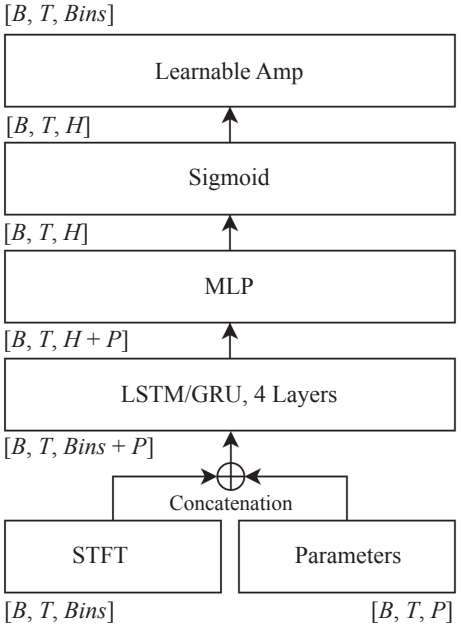

Fig. 2. Single-band RNN architecture. *B*, *T*, *Bins*, *P*, *H* represent the batch size, sequence length, number of STFT bins, number of conditioning parameters, and hidden size, respectively

The spectral amplification mask is formed by passing the MLP outputs through a sigmoid function and scaling the result with a learnable parameter, enabling both attenuation and amplification of individual frequency bins. The inverse transform is then applied to the modified magnitude components together with the original phase, resulting in the processed audio signal.

Multi-band LSTM

Real-time and causal processing is crucial for integrating deep learning models into DAWs, requiring models to operate efficiently across multiple channel configurations. For the widely used RNN architectures, LSTM and GRU, the computational complexity per time step of LSTM cells is given by $O(4(d + h)h)$ and for GRU cells by $O(3(d + h)h)$ where d is the input dimensionality, and h is the hidden state dimension. When d and h are of comparable magnitude, the total cost of the matrix-vector multiplications scales as $O(dh + h^2)$. Considering $d = \Theta(h)$ in our study, the resulting complexity is $O(h^2)$ with the input-to-hidden and hidden-to-hidden products contributing comparable amounts.

We propose a multi-band RNN architecture that mitigates computational complexity by partitioning the STFT magnitude bins into n equal frequency bands and enabling parallel processing across bands. In our approach, each frequency band is processed by an independent RNN branch with a reduced hidden state dimension:

$$d_b = \frac{d}{n}, \quad h_b = \frac{h}{n}.$$

By applying band-limited parameters to each branch, the computational complexity per time step of an LSTM cell is expressed as follows:

$$O(4(d_b + h_b)h_b) = O\left(4\frac{dh + h^2}{n^2}\right).$$

Considering the previously set $d = \Theta(h)$, the resulting term scales as $\frac{h^2}{n^2}$ and summing over all n parallel branches therefore yields:

$$O\left(n\frac{h^2}{n^2}\right) = O\left(\frac{h^2}{n}\right),$$

which represents an n-fold reduction relative to the single-band baseline with complexity of $O(h^2)$. Finally, the output of each branch is independently modulated using a sigmoid activation function and scaled by a trainable amplification coefficient. This design enables the network to adaptively scale the contributions of each band, resulting in a more efficient configuration (Fig. 3).

Datasets

Source audio from the Freesound [25] and Voice Cloning ToolKit (VCTK) [26] datasets was used to

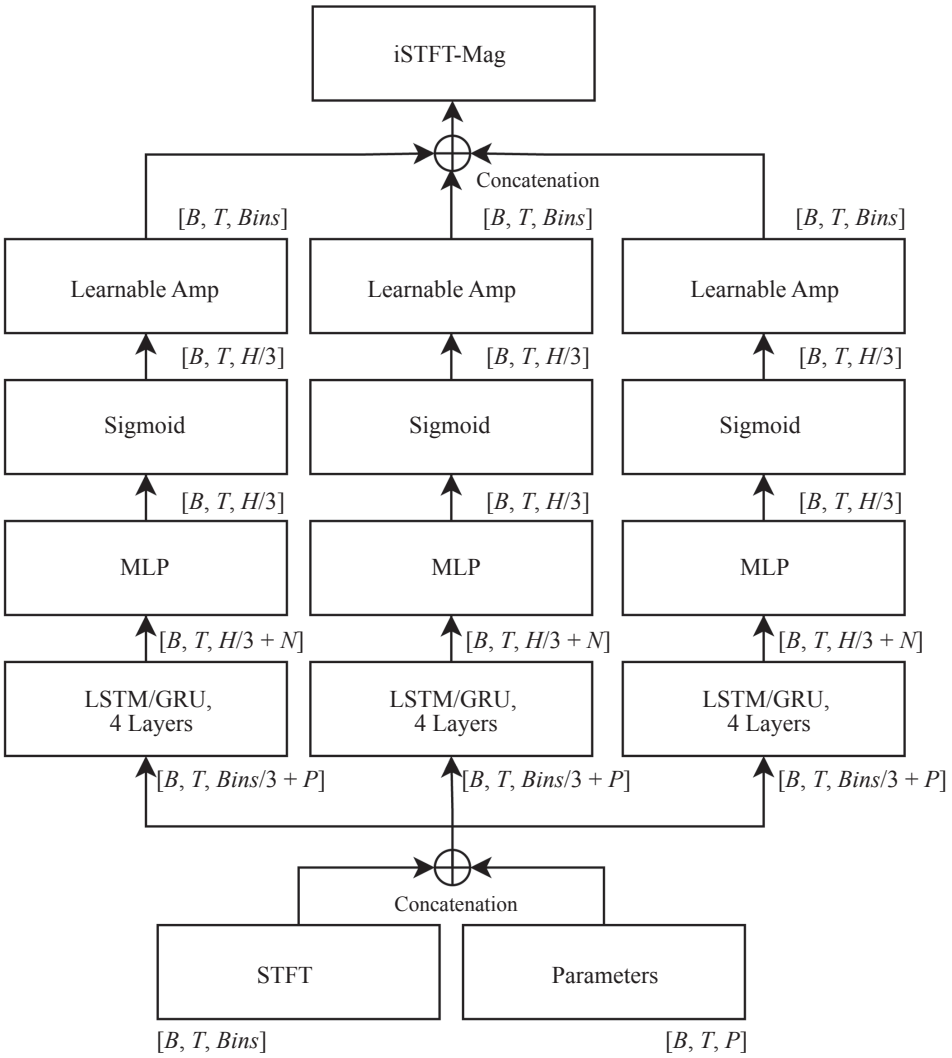

Fig. 3. Proposed multi-band RNN architecture. *B*, *T*, *Bins*, *P*, *H*, *N* represent the batch size, sequence length, number of STFT bins, number of conditioning parameters, hidden size, and dimensionality of conditioning vector respectively

construct the target dataset of processed outputs from the selected devices. As the model is intended for both music and voice processing tasks, the selected audio samples were carefully chosen to encompass a diverse range of audio categories. Specifically, we included complete music tracks, individual instrument recordings, and drum kit samples from the Freesound dataset as well as voice recordings captured via two distinct microphone setups from the VCTK dataset. All audio signals were concatenated into a single file and standardized to stereo 44.1 kHz, 16-bit Waveform Audio File (WAV) format. This concatenated file includes a 3-second silent interval between each source audio segment, ensuring that the compressor fully resets its attack and release stages. The resulting consolidated audio file served as the input audio for the collection of the device dataset.

For this study, we selected the consumer-grade analog compressor Alesis 3630 as the primary device for experimental evaluation. Additionally, we collected audio from discoDSP NightShine, a digital emulation of the Alesis 3630 available in VST plugin format, to better assess the performance and generalization of the model, while maintaining relevance to the original hardware. Audio signals were recorded using an Audient iD14 audio interface, introducing the influence of the interface analog-to-digital converter into the signal chain. Latency introduced by the audio interface was compensated by temporally aligning all output signals with their corresponding inputs using a cross-correlation method. We generated 144 unique parameter combinations using the Cartesian product of a predefined set and enhanced model generalization by adding randomly sampled combinations outside this set.

Although the datasets contain stereo recordings, the analysis focuses on single-channel modeling, using only the first channel of each pair as both the input and target. During training, audio segments of 1.5-second duration were randomly sampled as individual examples, while for evaluation, the segment length was extended to 4 seconds. The model state was not retained between audio segments during training or evaluation phases. To unify the amplitude ranges of the conditioning parameters within a single vector, parameters were normalized as follows: the ratio parameter is divided by 10, threshold by 10^2, attack and release by 10^3. For augmentation, we apply random phase inversion to both input and target signals with a probability $p = 0.5$.

Experiments

All models were trained for 120 epochs with an initial learning rate of $3 \cdot 10^{-4}$, using a Cosine Annealing learning rate scheduler and the Adam optimizer with default hyperparameters from the PyTorch framework. The training and evaluation were performed with mini-batches of 64 and 16 examples, respectively. All experiments were conducted on an NVIDIA RTX 3060 GPU with mixed precision enabled, with a total training duration of approximately 16 hours across all experiments. The training objective was formulated as the sum of the STFT and L_1 losses, where L_1 denotes the time-domain mean absolute error between the predicted and reference signals. In the following, X represents the magnitude spectrogram of the predicted signal and Y the magnitude spectrogram of the reference signal, where $\|\cdot\|_F$ denotes the Frobenius norm and $\|\cdot\|_1$ denotes the L_1 element-wise norm:

$$L_{STFT} = \frac{\| |X| - |Y| \|_F}{\| |Y| \|_F} + \| \ln|X| - \ln|Y| \|_1,$$

$$L = L_{STFT} + L_1.$$

Model performance was evaluated using four objective metrics. Both the L_1 and STFT losses, used during training, are included in the evaluation. The Root Mean Square (RMS) error is used to measure volume discrepancies. Lastly, we include the absolute LUFS (Loudness Units Full Scale) error between the predicted and target signals, computed using the ITU-R BS.1770 standard. These metrics are reported in the comparison tables in the following sections, with RMS metrics omitted in some cases for readability.

Computational complexity is reported in GMACs, corresponding to the number of billions of Multiply–Accumulate (MAC) operations per forward pass. Additionally, we assessed practical computational performance using the Real-Time (RT) factor, a metric representing the actual inference speed of the pipeline on consumer-grade hardware. The RT factor measurements were conducted on an Intel Core i5-12600KF CPU with chunk duration equal to 1.5 seconds. If P denotes the number of processed samples, T the processing time in seconds, and S represents the sampling rate, the RT factor is calculated using the following formula:

$$RT\,factor = \frac{P}{TS}.$$

Results

In this section, we present and analyze results for both the Alesis 3630 and discoDSP NightShine datasets in a unified table, facilitating a direct comparative assessment. The experimental setup and parameters were preserved across both datasets to ensure comparability.

Features

An important contribution of this work is the use of STFT-based feature representations combined with spectral amplification. We evaluated three configurations of window and hop sizes that satisfy the NOLA condition ordered by computational cost in GMACs. For this study, we used a one-layer, four-band LSTM/GRU network with a hidden state size of 1,024; and a 3-layer MLP that transforms hidden states to the required window size for the iSTFT transform.

Table 1 shows that the proposed configuration with smaller window and hop sizes enhances modeling quality for both devices. Among all evaluated settings, the best performance is achieved with a window size of 512 and a hop size of 256. We argue that variations in window size have minimal impact, while reductions in hop size substantially increase computational cost and influence

Table 1. Comparison of quality metrics and computational cost for different STFT configurations on the NightShine and Alesis 3630 datasets. Boldface indicates the best value in each column. "Params" column indicates the number of trainable parameters (M = million). Window and hop sizes are reported in samples

Model	Window	Hop	GMACs	Params	L_1	STFT	LUFS
discoDSP NightShine							
LSTM	2,048	1,024	0.21	3.2M	$7.72 \cdot 10^{-3}$	0.364	0.773
LSTM	1,024	512	0.33	2.5M	$5.22 \cdot 10^{-3}$	0.301	0.519
LSTM	512	256	0.57	2.2M	$4.97 \cdot 10^{-3}$	0.280	0.490
GRU	2,048	1,024	**0.17**	2.7M	$7.39 \cdot 10^{-3}$	0.335	0.693
GRU	1,024	512	0.28	2.1M	$5.53 \cdot 10^{-3}$	0.305	0.523
GRU	512	256	0.49	**1.9M**	$\mathbf{3.41 \cdot 10^{-3}}$	**0.266**	**0.333**
Alesis 3630							
LSTM	2,048	1,024	0.21	3.2M	$1.24 \cdot 10^{-2}$	0.670	1.082
LSTM	1,024	512	0.33	2.5M	$9.87 \cdot 10^{-3}$	0.527	0.694
LSTM	512	256	0.57	2.2M	$8.55 \cdot 10^{-3}$	0.472	0.486
GRU	2,048	1,024	**0.17**	2.7M	$1.08 \cdot 10^{-2}$	0.582	0.981
GRU	1,024	512	0.28	2.1M	$8.79 \cdot 10^{-3}$	0.506	0.498
GRU	512	256	0.49	**1.9M**	$\mathbf{7.98 \cdot 10^{-3}}$	**0.455**	**0.415**

model performance by producing more frames with finer temporal resolution.

Multi-band configurations

The impact of different band configurations on model performance was evaluated using a multi-band architecture with 1, 2, 4, and 8 bands. Feature extraction was performed using an STFT with a window size of 1,024 and a hop size of 512. Each configuration utilized a single-layer LSTM or GRU with a hidden state size of 1,024, followed by a 3-layer MLP.

As demonstrated in Table 2, increasing the number of bands effectively reduces computational complexity, including fewer GMAC operations, model parameters, and improved RT factors. However, this computational efficiency comes with a trade-off in audio-quality metrics. Specifically, the LSTM architecture achieves an average 12.1-fold improvement in RT factor when increasing from 1 to 8 bands, with a 29.3 % average relative degradation in the STFT loss. In comparison, the GRU architecture demonstrates a smaller real-time improvement of 3.1-fold while incurring a larger average STFT loss degradation of approximately 56 %.

Furthermore, emulating the Alesis 3630 processing resulted in a higher modeling error, as reflected by significantly higher L_1, STFT, and RMS losses compared to discoDSP NightShine. Notably, discrepancies between theoretical GMAC complexity and practical RT factor suggest hardware-specific optimizations significantly impact performance, particularly the lack of internal optimization negatively affecting GRU models configured with 8 bands.

Temporal parameters

The stateless inference setup, operating on approximately 1.5-second audio segments, motivates evaluation of model performance across different temporal parameter configurations of the audio compressor. In this evaluation, the compressor threshold was fixed at −10 dB and the ratio at 4:1, and the STFT performance of the 4-band GRU model was measured across a range of attack and release values assessed separately for each device (Fig. 4).

These plots indicate that an attack time of 0.1 ms and a release time ranging from 50 to 500 ms represent the most challenging conditions for the proposed model. This observation aligns with prior findings from [22], where fast attack dynamics similarly degraded model performance when emulating the CL-1B compressor using LSTM-ED architecture. We hypothesize that these limitations arise from the large window and hop sizes which extract coarse-grained features and thus fail to capture rapid transient changes induced by fast attack parameters.

Comparison

We compared our single-band and multi-band GRU and LSTM models with previously proposed raw waveform-based TCN and LSTM architectures from [21]. Specifically, causal variants of uTCN-100 and uTCN-300 were selected as baseline models. Because the waveform-based LSTM-32 operates on individual samples and the proposed model on blocks, their parameter counts are not directly comparable (Table 3).

The single-band GRU consistently outperforms waveform-based baselines. On NightShine, it reduces

Table 2. Comparison of quality metrics and computational cost for models with different band configurations on the NightShine and Alesis 3630 datasets. Boldface indicates the best value in each column. “Params” column indicates the number of trainable parameters (M = million)

Model	Bands	GMACs	Params	L_1	STFT	RMS	LUFS	RT Factor
discoDSP NightShine								
LSTM	1	1.29	10M	$4.98\cdot10^{-3}$	0.273	0.0093	**0.469**	13.4
LSTM	2	0.65	5.0M	$6.13\cdot10^{-3}$	0.299	0.0109	0.625	15.5
LSTM	4	0.33	2.5M	$5.36\cdot10^{-3}$	0.304	0.0096	0.569	53.9
LSTM	8	0.17	1.3M	$4.86\cdot10^{-3}$	0.324	0.0086	0.483	**162.3**
GRU	1	1.09	8.4M	$\mathbf{4.68\cdot10^{-3}}$	**0.244**	**0.0082**	0.488	32.1
GRU	2	0.55	4.2M	$5.54\cdot10^{-3}$	0.291	0.0094	0.547	63.8
GRU	4	0.28	2.1M	$5.04\cdot10^{-3}$	0.300	0.0089	0.475	107.4
GRU	8	**0.14**	**1.1M**	$6.55\cdot10^{-3}$	0.343	0.0111	0.638	99.4
Alesis 3630								
LSTM	1	1.29	10M	$9.28\cdot10^{-3}$	0.421	0.0173	0.628	13.4
LSTM	2	0.65	5.0M	$9.37\cdot10^{-3}$	0.470	0.0170	0.594	15.5
LSTM	4	0.33	2.5M	$9.85\cdot10^{-3}$	0.574	0.0178	0.686	53.9
LSTM	8	0.17	1.3M	$9.29\cdot10^{-3}$	0.589	0.0170	0.623	**162.3**
GRU	1	1.09	8.4M	$8.86\cdot10^{-3}$	**0.336**	**0.0159**	**0.423**	32.1
GRU	2	0.55	4.2M	$\mathbf{8.82\cdot10^{-3}}$	0.422	**0.0159**	0.505	63.8
GRU	4	0.28	2.1M	$9.39\cdot10^{-3}$	0.502	0.0168	0.551	107.4
GRU	8	**0.14**	**1.1M**	$9.13\cdot10^{-3}$	0.576	0.0166	0.639	99.4

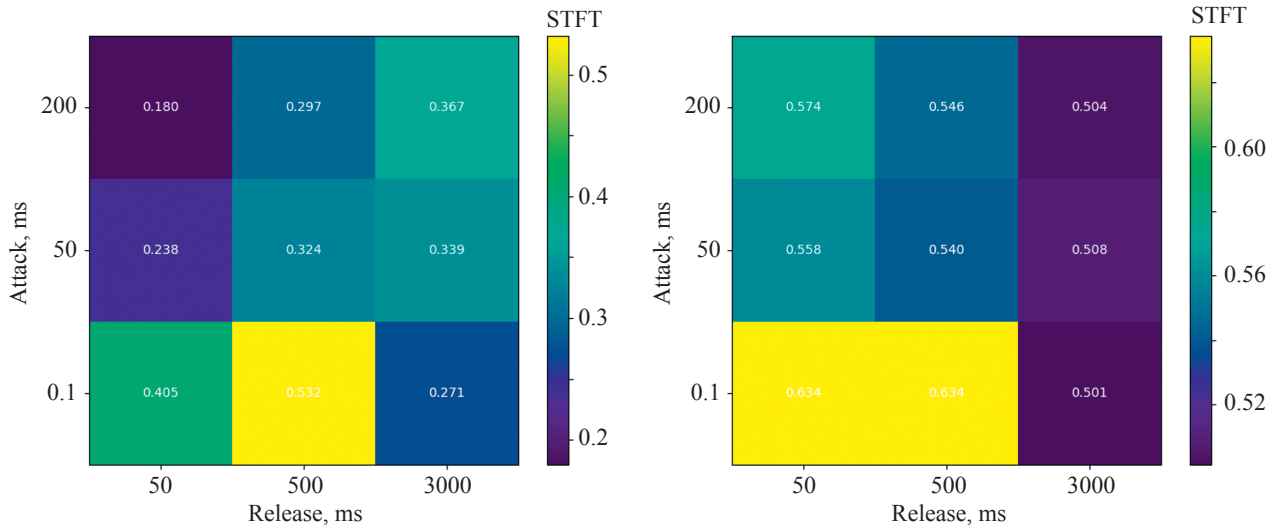

Fig. 4. Heatmaps of STFT-based losses for different attack and release configurations. Results are shown for the NightShine dataset (*a*) and the Alesis 3630 dataset (*b*)

L_1, RMS, and LUFS metrics by approximately 2.1–2.2× (and STFT loss by about 1.7×), while on the Alesis dataset it achieves reductions of roughly 1.3–1.6×, with LUFS improving by up to 3.5× in relative reduction. In addition, it provides an inference speed-up of approximately 35 % compared to uTCN-300, while maintaining comparable computational cost to uTCN-100. The proposed multi-band architectures substantially improved inference speeds, increasing RT factors from 13.4 to 53.9 for the LSTM and from 32.1 to 107.4 for the GRU. However, this computational advantage was accompanied by the degradation in objective metrics, mostly observed on the Alesis 3630 dataset. Notably, the STFT loss exhibited the most severe degradation, while the impact on L_1, RMS, and LUFS metrics was less significant.

Analysis of the results indicates that modeling the analog device behavior presents greater difficulty than modeling the digital emulation, as indicated by consistently higher error metrics across all evaluated architectures. Moreover, the most notable performance improvement among proposed models relative to the baselines was observed in the LUFS metric, with a two- to three-fold improvement, indicating improved accuracy in perceptual loudness modeling.

Table 3. Computational cost and quality metrics of the proposed models compared with the raw-audio-based baseline on the discoDSP NightShine and Alesis 3630 datasets. Boldface indicates the best value in each column. “Params” column indicates the number of trainable parameters (K = thousand, M = million)

Model	GMACs	Params	L_1	STFT	RMS	LUFS	RT Factor
discoDSP NightShine							
LSTM-32 [21]	0.35	**5K**	$1.44 \cdot 10^{-1}$	0.496	0.2276	0.924	32.2
uTCN-100 [21]	1.03	26K	$1.40 \cdot 10^{-1}$	0.501	0.2219	1.211	30.9
uTCN-300 [21]	2.48	51K	$9.90 \cdot 10^{-3}$	0.410	0.0179	1.076	23.7
LSTM-1B	1.29	10M	$4.98 \cdot 10^{-3}$	0.273	0.0093	**0.469**	13.4
GRU-1B	1.09	8.4M	$\mathbf{4.68 \cdot 10^{-3}}$	**0.244**	**0.0082**	0.488	32.1
LSTM-4B	0.33	2.5M	$5.36 \cdot 10^{-3}$	0.304	0.0096	0.569	53.9
GRU-4B	**0.28**	2.1M	$5.04 \cdot 10^{-3}$	0.300	0.0089	0.475	**107.4**
Alesis 3630							
LSTM-32 [21]	0.35	**5K**	$1.56 \cdot 10^{-2}$	0.862	0.0263	2.045	32.2
uTCN-100 [21]	1.03	26K	$9.68 \cdot 10^{-3}$	0.606	0.0192	1.715	30.9
uTCN-300 [21]	2.48	51K	$1.15 \cdot 10^{-2}$	0.525	0.0240	1.486	23.7
LSTM-1B	1.29	10M	$9.28 \cdot 10^{-3}$	0.421	0.0173	0.628	13.4
GRU-1B	1.09	8.4M	$\mathbf{8.86 \cdot 10^{-3}}$	**0.336**	**0.0159**	**0.423**	32.1
LSTM-4B	0.33	2.5M	$9.85 \cdot 10^{-3}$	0.574	0.0178	0.686	53.9
GRU-4B	**0.28**	2.1M	$9.39 \cdot 10^{-3}$	0.502	0.0168	0.551	**107.4**

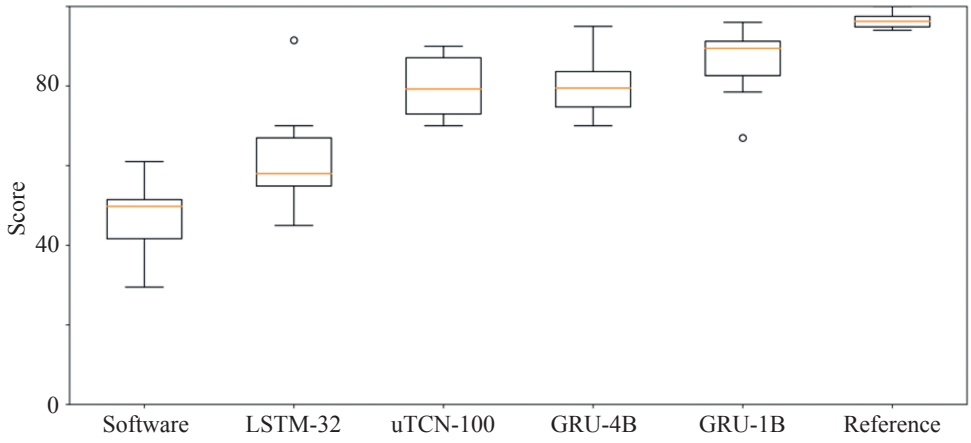

Fig. 5. Results of subjective listening study based on MUSHRA methodology

Subjective Evaluation

A subjective listening study was conducted using the MUSHRA protocol to complement the objective evaluation. Fig. 5 shows the aggregated results for the Alesis 3630 dataset presented as box plots. The test involved nine expert listeners with backgrounds in audio engineering and music production. One participant was excluded from the analysis for assigning a score below 90 to the reference condition, in accordance with MUSHRA guidelines. The listening test was performed in studio-grade environments, with each participant evaluating two stimuli per trial, resulting in a total of 18 ratings per condition. In addition to the neural models, the discoDSP NightShine plugin was included as a software-based emulation reference for the Alesis 3630 compressor.

The reference condition achieves the highest ratings, with a median close to the upper bound and minimal variance among the stimuli. GRU-1B achieves the highest perceptual scores among the neural models, ranking closest to the reference condition. GRU-4B and uTCN-100 show comparable performance, with GRU-4B showing a slightly higher median score and uTCN-100 demonstrating a narrower interquartile range. Importantly, GRU-4B reaches this level of perceptual quality while operating at approximately three times higher inference speed. The LSTM-32 model demonstrates moderate performance with lower median scores, while the software emulation receives the lowest ratings overall, revealing a clear perceptual gap relative to the neural approaches. Overall, these results indicate that GRU-1B provides the strongest subjective performance, while GRU-4B offers the most favorable accuracy–efficiency trade-off.

Conclusion

In this study, we introduced a Short-Time Fourier Transform (STFT)-based modeling approach that leverages magnitude response amplification and a spectral-based multi-band Recurrent Neural Network architecture that partitions the STFT magnitude spectrum into multiple parallel frequency bands for independent processing, achieving a balance between modeling precision and computational complexity. To support our experiments, we developed a dedicated data collection procedure and created two parallel datasets: recordings from the consumer-grade analog compressor Alesis 3630 and from discoDSP NightShine, its corresponding digital emulation.

Experimental results demonstrate that the proposed single-band and multi-band architectures consistently outperform raw waveform-based Long Short-Term Memory and Temporal Convolutional Network baselines across multiple objective audio-quality metrics and a subjective evaluation, while significantly reducing computational cost. However, a limitation of the current approach is its reliance on the original phase during reconstruction, which can produce phase dispersion artifacts. Methods that jointly model amplitude and phase therefore represent an important direction for future research. The present validation is limited to two closely related devices, and further work is needed to assess the generalizability of the method to compressors with different architectures. Future research will also address adaptive temporal resolution, model compression, and perceptual loss integration to improve performance and audio quality[1].

[1] Code and datasets are available at: https://github.com/4antii/vca-comp (accessed: 10.03.2026).

References

1. Wilmering T., Moffat D., Milo A., Sandler M. A history of audio effects. *Applied Sciences*, 2020, vol. 10, no. 3, pp. 791. https://doi.org/10.3390/app10030791
2. Montenegro J. Design of an audio compressor with digital control. *TECCIENCIA*, 2021, vol. 16, no. 30, pp. 51–64. https://doi.org/10.18180/tecciencia.2021.30.4

Литература

1. Wilmering T., Moffat D., Milo A., Sandler M. A history of audio effects // Applied Sciences. 2020. V. 10. N 3. P. 791. https://doi.org/10.3390/app10030791
2. Montenegro J. Design of an audio compressor with digital control // TECCIENCIA. 2021. V. 16. N 30. P. 51–64. https://doi.org/10.18180/tecciencia.2021.30.4

3. Välimäki V., Reiss J. All about audio equalization: solutions and frontiers. *Applied Sciences*, 2016, vol. 6, no. 5, pp. 129. https://doi.org/10.3390/app6050129
4. Réveillac J.-M. *Musical Sound Effects: Analog and Digital Sound Processing*. Wiley-ISTE, 2017, 558 p.
5. Chowdhury J. A comparison of virtual analog modelling techniques for desktop and embedded implementations. *arXiv*, 2020. arXiv:2009.02833. https://doi.org/10.48550/arXiv.2009.02833
6. Purwins H., Li B., Virtanen T., Schlüter J., Chang S.-Y., Sainath T. Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, 2019, vol. 13, no. 2, pp. 206–219. https://doi.org/10.1109/jstsp.2019.2908700
7. Liu X., Sahidullah M., Kinnunen T. A comparative re-assessment of feature extractors for deep speaker embeddings. *Proc. of the Annual Conference of the International Speech Communication Association Interspeech*, 2020, pp. 3221–3225. https://doi.org/10.21437/interspeech.2020-1765
8. Sun Y., Yang L., Zhu H., Hao J. Funnel deep complex U-Net for phase-aware speech enhancement. *Proc. of the Annual Conference of the International Speech Communication Association Interspeech*, 2021, pp. 161–165. https://doi.org/10.21437/Interspeech.2021-10
9. Kong J., Kim J., Bae J. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Proc. of the 34th International Conference on Neural Information Processing Systems*, 2020, pp. 17022–17033.
10. Zölzer U. *DAFX: Digital Audio Effects*. Wiley, 2011, 624 p
11. Kates J.M. Principles of digital dynamic-range compression. *Trends in Amplification*, 2005, vol. 9, no. 2, pp. 45–76. https://doi.org/10.1177/108471380500900202
12. Giannoulis D., Massberg M., Reiss J.D. Digital dynamic range compressor design—A tutorial and analysis. *Journal of the Audio Engineering Society*, 2012, vol. 60, no. 6, pp. 399–408.
13. D'Angelo S. Lightweight virtual analog modeling. *Proc. of the 22nd Colloquio di Informatica Musicale* (CIM), 2018.
14. Eichas F., Zölzer U. Virtual analog modeling of guitar amplifiers with Wiener–Hammerstein models. *Proc.of the 44th Annual Convention on Acoustics* (DAGA), 2018.
15. Cheng C.M., Peng Z.K., Zhang W.M., Meng G. Volterra-series-based nonlinear system modeling and its engineering applications: A state-of-the-art review. *Mechanical Systems and Signal Processing*, 2017, vol. 87, part A. pp. 340–364. https://doi.org/10.1016/j.ymssp.2016.10.029
16. van den Oord A., Dieleman S., Zen H., Simonyan K., Vinyals O., Graves A., et al. WaveNet: A generative model for raw audio. *arXiv*, 2016. arXiv:1609.03499. https://doi.org/10.48550/arXiv.1609.03499
17. Wright A., Damskägg E.-P., Välimäki V. Real-time black-box modelling with recurrent neural networks. *Proc. of the 22nd International Conference on Digital Audio Effects* (DAFx-19), 2019, pp. 1–9.
18. Ramirez M.A.M., Benetos E., Reiss J.D. Deep learning for black-box modeling of audio effects. *Applied Sciences*, 2020, vol. 10, no. 2, pp. 638. https://doi.org/10.3390/app10020638
19. Damskägg E.-P., Juvela L., Välimäki V. Real-time modeling of audio distortion circuits with deep learning. *Proc. of the 16th Sound and Music Computing Conference*, 2019, pp. 332–339.
20. Hawley S.H., Colburn B., Mimilakis S.I. SignalTrain: profiling audio compressors with deep neural networks. *arXiv*, 2019. arXiv:1905.11928. https://doi.org/10.48550/arXiv.1905.11928
21. Steinmetz C.J., Reiss J.D. Efficient neural networks for real-time analog audio effect modeling. *arXiv*, 2021. arXiv:2102.06200. https://doi.org/10.48550/arXiv.2102.06200
22. Simionato R., Fasciani S. Fully conditioned and low-latency black-box modeling of analog compression. *Proc. of the International Conference on Digital Audio Effects Dafx*, 2023.
23. Yin H., Cheng G., Steinmetz C.J., Yuan R., Stern R.M., Dannenberg R.B. Modeling analog dynamic range compressors using deep learning and state-space models. *arXiv*, 2024. arXiv:2403.16331. https://doi.org/10.48550/arXiv.2403.16331
24. Simionato R., Fasciani S. Modeling time-variant responses of optical compressors with selective state space models. *AES Journal of the Audio Engineering Society*, 2025, vol. 73, no. 3. pp. 144–165. https://doi.org/10.17743/jaes.2022.0194
25. Fonseca E., Favory X., Pons J., Font F., Serra X. FSD50K: An open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022, vol. 30, pp. 829–852. https://doi.org/10.1109/taslp.2021.3133208

3. Välimäki V., Reiss J. All about audio equalization: solutions and frontiers // Applied Sciences. 2016. V. 6. N 5. P. 129. https://doi.org/10.3390/app6050129
4. Réveillac J.-M. Musical Sound Effects: Analog and Digital Sound Processing. Wiley-ISTE, 2017. 558 p.
5. Chowdhury J. A comparison of virtual analog modelling techniques for desktop and embedded implementations // arXiv. 2020. arXiv:2009.02833. https://doi.org/10.48550/arXiv.2009.02833
6. Purwins H., Li B., Virtanen T., Schlüter J., Chang S.-Y., Sainath T. Deep learning for audio signal processing // IEEE Journal of Selected Topics in Signal Processing. 2019. V. 13. N 2. P. 206–219. https://doi.org/10.1109/jstsp.2019.2908700
7. Liu X., Sahidullah M., Kinnunen T. A comparative re-assessment of feature extractors for deep speaker embeddings // Proc. of the Annual Conference of the International Speech Communication Association Interspeech. 2020. P. 3221–3225. https://doi.org/10.21437/interspeech.2020-1765
8. Sun Y., Yang L., Zhu H., Hao J. Funnel deep complex U-Net for phase-aware speech enhancement // Proc. of the Annual Conference of the International Speech Communication Association Interspeech. 2021. P. 161–165. https://doi.org/10.21437/Interspeech.2021-10
9. Kong J., Kim J., Bae J. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis // Proc. of the 34th International Conference on Neural Information Processing Systems. 2020. P. 17022–17033.
10. Zölzer U. DAFX: Digital Audio Effects. Wiley, 2011. 624 p.
11. Kates J.M. Principles of digital dynamic-range compression // Trends in Amplification. 2005. V. 9. N 2. P. 45–76. https://doi.org/10.1177/108471380500900202
12. Giannoulis D., Massberg M., Reiss J.D. Digital dynamic range compressor design—A tutorial and analysis // Journal of the Audio Engineering Society. 2012. V. 60. N 6. P. 399–408.
13. D'Angelo S. Lightweight virtual analog modeling // Proc. of the 22nd Colloquio di Informatica Musicale (CIM). 2018.
14. Eichas F., Zölzer U. Virtual analog modeling of guitar amplifiers with Wiener–Hammerstein models // Proc.of the 44th Annual Convention on Acoustics (DAGA). 2018.
15. Cheng C.M., Peng Z.K., Zhang W.M., Meng G. Volterra-series-based nonlinear system modeling and its engineering applications: A state-of-the-art review // Mechanical Systems and Signal Processing. 2017. V. 87. Part A. P. 340–364. https://doi.org/10.1016/j.ymssp.2016.10.029
16. van den Oord A., Dieleman S., Zen H., Simonyan K., Vinyals O., Graves A., et al. WaveNet: A generative model for raw audio // arXiv. 2016. arXiv:1609.03499. https://doi.org/10.48550/arXiv.1609.03499
17. Wright A., Damskägg E.-P., Välimäki V. Real-time black-box modelling with recurrent neural networks // Proc. of the 22nd International Conference on Digital Audio Effects (DAFx-19). 2019. P. 1–9.
18. Ramirez M.A.M., Benetos E., Reiss J.D. Deep learning for black-box modeling of audio effects // Applied Sciences. 2020. V. 10. N 2. P. 638. https://doi.org/10.3390/app10020638
19. Damskägg E.-P., Juvela L., Välimäki V. Real-time modeling of audio distortion circuits with deep learning // Proc. of the 16th Sound and Music Computing Conference. 2019. P. 332–339.
20. Hawley S.H., Colburn B., Mimilakis S.I. SignalTrain: profiling audio compressors with deep neural networks // arXiv. 2019. arXiv:1905.11928. https://doi.org/10.48550/arXiv.1905.11928
21. Steinmetz C.J., Reiss J.D. Efficient neural networks for real-time analog audio effect modeling // arXiv. 2021. arXiv:2102.06200. https://doi.org/10.48550/arXiv.2102.06200
22. Simionato R., Fasciani S. Fully conditioned and low-latency black-box modeling of analog compression // Proc. of the International Conference on Digital Audio Effects Dafx. 2023.
23. Yin H., Cheng G., Steinmetz C.J., Yuan R., Stern R.M., Dannenberg R.B. Modeling analog dynamic range compressors using deep learning and state-space models // arXiv. 2024. arXiv:2403.16331. https://doi.org/10.48550/arXiv.2403.16331
24. Simionato R., Fasciani S. Modeling time-variant responses of optical compressors with selective state space models // AES Journal of the Audio Engineering Society. 2025. V. 73. N 3. P. 144–165. https://doi.org/10.17743/jaes.2022.0194
25. Fonseca E., Favory X., Pons J., Font F., Serra X. FSD50K: An open dataset of human-labeled sound events // IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2022. V. 30. P. 829–852. https://doi.org/10.1109/taslp.2021.3133208

26. Yamagishi J., Veaux C., MacDonald K. CSTR VCTK corpus: English multi-speaker corpus for CSTR Voice Cloning Toolkit (version 0.92). *University of Edinburgh, Centre for Speech Technology Research* (CSTR), 2019, https://doi.org/10.7488/ds/2645

26. Yamagishi J., Veaux C., MacDonald K. CSTR VCTK corpus: English multi-speaker corpus for CSTR Voice Cloning Toolkit (version 0.92) // University of Edinburgh, Centre for Speech Technology Research (CSTR). 2019. https://doi.org/10.7488/ds/2645

Authors

Andrei F. Balykin — PhD Student, St. Petersburg State University (SPbU), Saint Petersburg, 199034, Russian Federation, SC 58548795200, https://orcid.org/0009-0003-1554-2873, st054659@student.spbu.ru

Ivan S. Blekanov — PhD, Associate Professor, St. Petersburg State University (SPbU), Saint Petersburg, 199034, Russian Federation, SC 56149559700, https://orcid.org/0000-0002-7305-1429, i.blekanov@spbu.ru

Авторы

Балыкин Андрей Федорович — аспирант, Санкт-Петербургский государственный университет, Санкт-Петербург, 199034, Российская Федерация, SC 58548795200, https://orcid.org/0009-0003-1554-2873, st054659@student.spbu.ru

Блеканов Иван Станиславович — кандидат технических наук, доцент, заведующий кафедрой, Санкт-Петербургский государственный университет, Санкт-Петербург, 199034, Российская Федерация, SC 56149559700, https://orcid.org/0000-0002-7305-1429, i.blekanov@spbu.ru

Received 22.09.2025
Approved after reviewing 18.02.2026
Accepted 18.03.2026

Статья поступила в редакцию 22.09.2025
Одобрена после рецензирования 18.02.2026
Принята к печати 18.03.2026