CONMOD: CONTROLLABLE NEURAL FRAME-BASED MODULATION EFFECTS

Gyubin Lee

School of Computing
KAIST
Daejeon, South Korea
gbstorm81@kaist.ac.kr

Hounsu Kim

Graduate School of Culture Technology
KAIST
Daejeon, South Korea
hanshounsu@kaist.ac.kr

Junwon Lee

Kim Jaechul Graduate School of AI
KAIST
Daejeon, South Korea
james39@kaist.ac.kr

Juhan Nam

Graduate School of Culture Technology
KAIST
Daejeon, South Korea
juhan.nam@kaist.ac.kr

ABSTRACT

Deep learning models have seen widespread use in modelling LFO-driven audio effects, such as phaser and flanger. Although existing neural architectures exhibit high-quality emulation of individual effects, they do not possess the capability to manipulate the output via control parameters. To address this issue, we introduce **C**ontrollable **N**eural Frame-based **Mod**ulation Effects (CONMOD), a single black-box model which emulates various LFO-driven effects in a frame-wise manner, offering control over LFO frequency and feedback parameters. Additionally, the model is capable of learning the continuous embedding space of two distinct phaser effects, enabling us to steer between effects and achieve creative outputs. Our model outperforms previous work while possessing both controllability and universality, presenting opportunities to enhance creativity in modern LFO-driven audio effects. Additional demo of our model is available in the accompanying website.[1]

1. INTRODUCTION

The study of virtual analog (VA) modelling via deep neural networks have become a well-established topic nowadays. Neural networks are strongly capable of modelling the behavior of the system in a black-box manner, and is being applied in various VA modelling tasks [1, 2, 3, 4]. Furthermore, the strongpoint of deep neural networks as a regression model anticipates an intuitive and creative effect by adjusting a representative embedding vector of the trained model [5, 6].

The field of VA is largely divided into white-box, grey-box and black-box approaches. The white-box approach discretizes the acoustic system of the analog circuit structure to obtain a digital model [7, 8]. The black-box approach incorporates parametric models which emulate the input-output behavior of the analog circuits without knowing the physical structure [9, 10, 11]. Grey-box

---

approach leverages some knowledge of the acoustic system to better emulate the parametric model [12, 13]. Among them, the black-box approach has recently drawn a great attention as it can behave more robust and general in the case of deep neural networks when the model is well-structured and trained with a sufficient amount of data [14, 11, 15, 16, 2, 3].

Phaser and flanger effects modulate the input audio using a periodic signal typically within the human preferred range (0.1-10Hz), commonly referred to as LFO-driven effects. Core aspect of modelling these effects is maintaining precise periodicity regardless of the input audio duration. Therefore, early neural architectures integrated an LFO signal in the form of a rectified sine wave [17, 18] or a sinusoidal wave [4] to enable the model to focus on generating the spectral structure. Early works related to these approach [17, 18] extracts the LFO signal manually, which is a clear disadvantage for training. [19] further proposed a CNN-based LFO extraction model, achieving plausible results for unseen audio and effects. Unlike existing black-box approaches that learn the direct mapping from dry input audio sets to corresponding wet output audio sets, Carson et al. simplifies the role of model into emulating solely the effect of a phaser pedal using a frame-based processing [4]. Nevertheless, the paper does not explore the model's controllability with respect to control parameters(e.g. LFO frequency, feedback, etc), even though directly manipulating the sinusoidal wave may alter the LFO frequency. And even so, the model is controllable with only a single control parameter.

In this paper, we build upon the model architecture proposed by Carson et al. [4] and introduce a novel controllable neural effects designed for two types of modulation effects, phaser and flanger. Specifically, our neural architecture enables adjustment of the output with user-specified rate and feedback parameters for the target effect. Rather than directly optimizing the parameters related to the physics of the phaser effect, we optimize a neural architecture consisting of long-term short-memory (LSTM), multi-layer perceptron (MLP), and conditional feature wise linear modulation (FiLM) blocks[20]. To enhance the model's controllability on the LFO frequency, we introduce a training procedure which optimizes multiple LFO signals within a single neural network. We compare our work with a baseline model [4] and show that our model shows superior accuracy in trained control parameter settings, while also exhibiting impressive performance with unseen control parameters. Inspired by [5, 6], we further explored our

model's capability to train multiple modulation effects simultaneously. Through training on two unique phaser effects concurrently, and by manipulating each effect's learned embedding, we show that our model can produce intriguing mixed results between the two distinct phaser effects. The main contributions of our work can be summarized as follows:

- We introduce a controllable neural frame-based modulation effect pedal targeting phaser and flanger, which outperforms previous model in emulation capability.

- Our model possesses robust controllability on multiple control parameters, capable of modelling effects on unseen control parameter settings, a feat unattainable by previous models.

- For the first time, we train a single model on two distinct phaser effects and demonstrate the model's steerability by manipulating the learned representative embeddings.

2. TARGET MODULATION EFFECTS

In this section, we briefly explain the characteristics of our target modulation effects, and the specific effect pedals that were trained on our model. For each of the effects, we trained our model with digital and analog devices.

2.1. Phaser effects and its controls

The phaser pedal modulates the input signal by utilizing low frequency oscillating notch filters or low-order all-pass filters. When cascaded, these filters periodically attenuate certain frequency levels of the signal, thereby introducing soft, moving sounds. In this paper, we focus on learning to control the two main parameters: LFO frequency and feedback, following [4]. LFO frequency, also called *Speed* or *Rate*, governs the frequency of the LFO signal, while feedback adjusts the amount of filtered output signal that is fed back to the input. The adjustment of feedback leads to an overall change in the shape of the frequency response, notably bringing resonance peaks in the unattenuated area [21]. Feedback is often termed *Color* due to its ability to alter the frequency response shape.

2.1.1. Studio One Phaser[2]

Studio One Phaser[2] is a built-in digital phaser included in Studio One, featuring various controllable parameters such as Speed, Feedback, Center, and Range [22]. In Phaser[2] the *Speed* knob, which controls the LFO frequency, allows for continuous adjustment within a range of 0 Hz to 10 Hz. The *Feedback* knob also offers continuous range from 0 to 95%. For the remaining controls, we maintain the default values.

2.1.2. Moore Liquid Phaser

Moore Liquid Phaser is a modern analog phaser offering selection among 5 distinct *Phaser Type*s and featuring controls among 3 different *Waveform Type*s(Round, Triangular, or Square), *Speed*, and *Color*. We set the *Waveform Type* to Round, and the *Phaser Type* to feedback mode, which corresponds to the 12 o' clock direction of the knob ('III').

2.2. Flanger effects and its controls

The flanger effect has a similar effect compared to the phaser effect, generating spectral notches. The main difference is that flanger creates notches by introducing a variable-length feedforward delayline which leads to a comb filter. Therefore the notches, in general, are regularly distributed across the frequency spectrum, extending infinitely. In this work, we adopt a similar approach as in the phaser case, concentrating on the LFO frequency and feedback parameters, which function on the same principle as in the phaser effect.

2.2.1. Studio One Flanger

Studio One Flanger is a built-in digital flanger with controllable parameters similar to those found in a phaser, such as *Speed*, *Feedback*, *Delay*, and *LFO amount*. The *Speed* control allows for adjustments within a continuous value from 0.01Hz to 10Hz. The *Feedback* knob provides a range from -90% to 90%, where negative values lead to inverted feedback [22].

2.2.2. Flamma FC15 Analog Flanger

The Flamma FC 15 analog flanger is a classic device carrying two effect modes, namely filter and normal, along with controllable knobs for *Color*, *Range*, and *Rate* [23]. We maintain the *Modes* setting at normal, and set the *Range* parameter to the 12 o' clock position, while training our model on variable settings for *Color* and *Rate*.

3. METHODOLOGY

We introduce a novel black-box methodology controllable via various effect parameters, which we refer to as CONMOD(**Co**ntrollable **N**eural Frame-based **Mod**ulation Effects). Our model efficiently models the audio effect frame-wise with less reliance on prior domain knowledge.

3.1. Model

Our model predicts the frame-dependent transfer function of the target effect using a sinusoidal LFO signal with trainable frequency and initial phase as input, following [4]. This approach approximates the effects, which are analogous to an IIR filter, by representing them as a truncated impulse response, where the truncation length is equal to the window size of each frame. As the frame-based process remains differentiable throughout the system, we can train the model to predict solely the modulation effect, with the training objective being the target wet audio waveform. The input for the training phase is a chirp signal, and the target pair is the wet chirp signal.

The main difference with [4] lies in the neural network structure responsible for predicting the transfer function. We begin by employing LSTM on the LFO signal to enhance the model's robustness under varying LFO frequency conditions. The features are then passed through fully-connected (FC) layers to shape the frame-dependent transfer function. FiLM layers are inserted between these FC layers to incorporate the feedback condition and effect type embeddings. The final output comprises predicted transfer functions for each input frame, resulting in a complex matrix equivalent in size to the STFT matrix of the dry signal. After element-wise multiplication of these matrices, an inverse STFT is
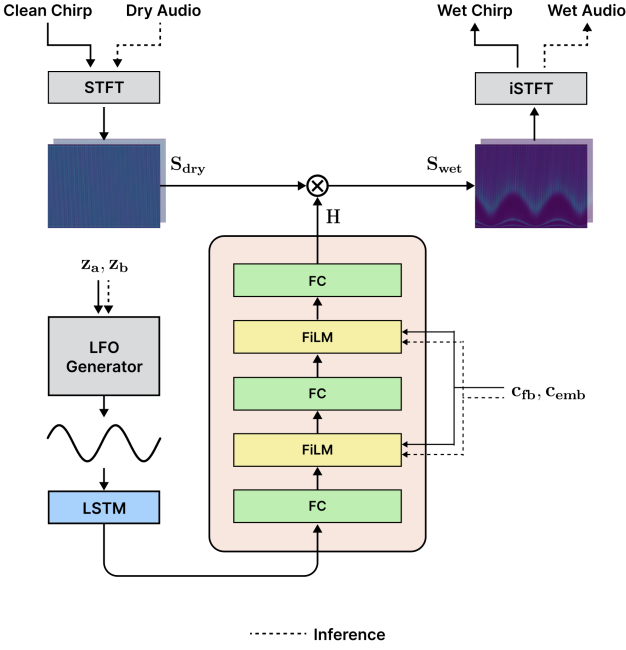
< **10** >

Figure 1: Overall model architecture. Blocks that have both dotted lines and solid lines as input indicate that the input data is distinctive for the training phase and inference phase. Grey-colored blocks indicate they have no trainable parameters and their processes are differentiable, thus enabling backpropagation. z_a, z_b, and c_{emb} are trainable parameters.

conducted to derive the predicted wet signal. The comprehensive model architecture is depicted in Figure 1.

Our model employs both time domain loss and frequency domain loss for the training objective. For the time domain loss, we utilize the error-to-signal ratio (ESR) defined as:

$$\mathcal{L}_{ESR}(y, \hat{y}) = \frac{\sum_{l=0}^{L-1}(y[l] - \hat{y}[l])^2}{\sum_{l=0}^{L-1} y[l]^2} \quad (1)$$

where L is the length of the training data in samples. For the frequency domain loss, we adopt the multi-resolution spectral loss (MRSL) defined as:

$$\mathcal{L}_{MRSL} = \sum_i ||S_i - \hat{S}_i||_1 + ||\log S_i - \log \hat{S}_i||_1 \quad (2)$$

where i is the index of each FFT size, which is set as (512, 1024, 2048). The neighboring frames during the STFT are overlapped by 75%. Consequently, the final training objective of our model becomes:

$$\mathcal{L}(y, \hat{y}) = \lambda \mathcal{L}_{ESR}(y, \hat{y}) + \mathcal{L}_{MRSL} \quad (3)$$

where the weight λ is set to 2 throughout all experiments.

3.2. Conditioning the control parameters

We employed the FiLM layer to apply the feedback condition to the model [20]. This is mathematically formalized as the following:

$$\text{MLP}([c_{fb}, c_{emb}]) = [\gamma, \beta]$$
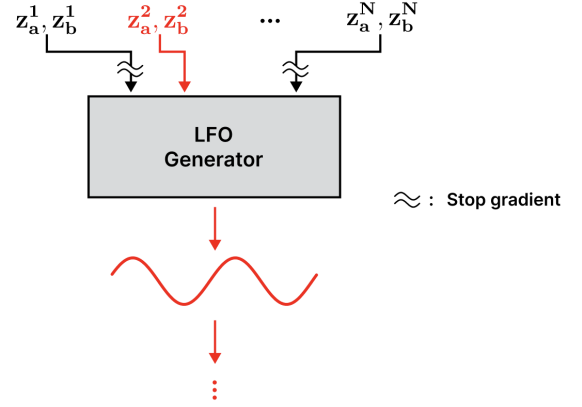$$\text{FiLM: } y_{fc} = \gamma \cdot x_{fc} + \beta \quad (4)$$

Figure 2: Proposed model training technique on multiple LFO frequency controls. For a random input-output pair with specified LFO frequency settings, only the corresponding z_a, z_b parameters are optimized. Figure depicts the case when z_a^2, z_b^2 are optimized. Other parameters are not updated by employing stop gradient operation.

where c_{fb} denotes the feedback, c_{emb} is the embedding of effect type, $[\cdot, \cdot]$ denotes simple concatenation of two vectors, γ, β are the output factors of FiLM, and x_{fc}, y_{fc} are the intermediate input feature and its affine transformed output respectively. Effect type embedding c_{emb} is only used when the model is trained on multiple phaser effects. (refer to Section 5.4).

In [4], the time-varying behavior of the target wet audio is modeled as a sinusoidal LFO signal and is fitted through gradient descent during the training phase [24]. The sinusoidal LFO signal is generated from input parameters directly related to the frequency z_a and initial phase z_b, which is then fed into the LSTM block. In our approach, we follow the LFO fitting methodology in [4] and additionally train the model on target wet audio modulated with multiple LFO frequencies to achieve robust controllability over frequency. The parameters being optimized are thus $(z_a^1, z_b^1), (z_a^2, z_b^2), ..., (z_a^N, z_b^N)$ where each of the N couple of parameters is assigned to train different LFO signals. When an audio pair with LFO frequency corresponding to the i-th signal is input to the model, other parameters are not updated, utilizing the stop gradient operator [25], as illustrated in Figure 2.

4. EXPERIMENTS

4.1. Dataset Preparation

We utilize a chirp-train signal with impulse period 40ms, passed through cascaded 64 all-pass filters with a coefficient of $p = 0.9$, following [4]. Training data consists of 6.67 seconds of chirp signal, while direct-input (DI) recordings provided in [4][2] are utilized for inference. For digital phaser and flanger effects, we vary the LFO frequency(Hz) among 0.23, 0.73, 1.13 and the feedback (%) parameters among 0, 25, 50, 75. The number of control parameter values are selected heuristically, resulting in a total of 12 audio samples used for training the model. All other effect parameters except for LFO frequency and feedback parameter remain fixed. Overall parameter settings are organized in Table 1 Notably, 23,

[2]https://github.com/a-carson/ddsp-phaser/tree/main/audio_data

<   **11**   >

73, and 113 are all prime numbers, ensuring that the LFO generator provides various and non-overlapping LFO phase to the LSTM block.

Table 1: *Control parameter experiment settings for the digital effect.*

Studio One Digital Phaser (Phaser^2)		
Modulation	Stages	6
	Speed (Hz)	0.23, 0.73, 1.13
Color	Center	1300
	Range	80%
	Feedback (%)	0, 25, 50, 75
Global	Stereo Spread	2.0%
	Mix	70.0%
Studio One Digital Flanger		
Color	LFO-Amount	80%
	Speed (Hz)	0.23, 0.73, 1.13
Modulation	Feedback (%)	0, 25, 50, 75
	Delay	1.50ms
Global	Mix	70.0%

The LFO frequency of the analog effect was adjusted to closely match the LFO settings of the digital effects. For the analog phaser, the LFO frequency knob was set to 8 o'clock, halfway between 8 o'clock and 9 o'clock, and 9 o'clock, resulting in corresponding LFO frequencies (in Hz) of 0.3549, 0.7519, 1.430. Similarly, for the analog flanger, the LFO frequency knob was set to 12 o'clock, 3 o'clock, and halfway between 3 o'clock and 4 o'clock, resulting in corresponding LFO frequencies of 0.2438, 0.6816, 1.004. Regarding the feedback parameter, we determined the maximum feedback value that produced an audible sound (3 o'clock for the analog phaser and 12 o'clock for the analog flanger). Subsequently, we set four feedback values that evenly divided the range between no feedback and the maximum audible feedback. We designated each feedback such that a clockwise rotation of the knob by $45°$ corresponds to a 12.5 increment in feedback. Consequently, the resulting feedback labels are 0.0, 12.5, 37.5, 62.5 for the analog phaser and 0.0, 12.5, 25, 37.5 for the analog flanger. As with the digital effects, all other effect parameters except for the LFO frequency and feedback parameters, are held constant. Specific parameter settings are presented in Table 2.

Table 2: *Control parameter experiment settings for the analog effect.*

Moore Liquid Phaser	
Phaser Type	12 o' clock
LFO shape	Round
Speed (Hz)	0.3549, 0.7519, 1.430
Color	0.0, 12.5, 37.5, 62.5
Flamma FC15 Analog Flanger	
Mode	Normal
Range	12 o'clock
Rate (Hz)	0.2438, 0.6816, 1.004
Color	0.0, 12.5, 25, 37.5

4.2. Baselines

For model comparison, we choose the previous grey-box phaser model[4] as our baseline. This model has a frame-based approach similar to our proposed model but is limited in learning the phaser effect due to its confined structure. Consequently, it cannot be trained for other time-varying filters or delay-based audio effects such as flanger. Additionally, although the baseline model may have the potential for controllability by manipulating the LFO signal, this capability is not explicitly explored in the paper. Therefore, multiple baseline models are trained individually for each LFO frequency and feedback condition. We conducted the comparison for both digital and analog phasers.

4.3. Training and Inference Setup

To ensure that all LFO-related parameters are learned independently, all models are trained with a batch size of 1. We train audio with sample rate of 44.1kHz and adopt an FFT size of 4096, hop size of 441 and a fixed frame size of 1764 throughout the experiments. Adam optimizer with exponential learning rate decay ($\gamma = 0.9997$) and an initial learning rate of 0.001 is employed. The models are trained for 10,000 epochs while each epoch is composed of 12 chirp signals, resulting in a total training duration of 80.04 seconds. The hidden size of LSTM layer and MLP layers are 32 and 512 respectively, where a single MLP layer predicts the real and imaginary part of the STFT matrix. With these settings, the model has total 2.4M learnable parameters. For the digital effect, the initial LFO phase is identical for each recording. Consequently, only three LFO generator parameters, corresponding to the variations in LFO frequency, need to be trained. In the case of analog effect, we could not designate a desired initial LFO phase during the recording of the 12 wet targets. Therefore the model had to predict 12 distinct LFO signals for each of the 12 wet chirp signals, leading to unstable convergence. Since we cannot obtain the groundtruth initial phase and LFO frequency values for the analog effect, we conducted pilot experiments by training the 12 signals individually and provide the model the trained $\{z_a^i\}_{i=1}^{12}$, $\{z_b^i\}_{i=1}^{12}$ as initial values to stabilize the training.

For quantitative evaluation, we utilized the ESR metric on 10s direct-input(DI) guitar recording.

5. RESULTS

In this section, we aim to evaluate the performance of our model in modelling digital and analog phaser/flanger effects compared to the baseline model. We also highlight our model's controllability across various control parameter settings. Ablation studies on long-term audio inference capability and the effect of training on multiple LFO frequencies are conducted. Lastly, we showcase a model trained on two distinct phasers simultaneously, which implies the potential of our approach in universal modelling.

5.1. Main Results

5.1.1. Digital Phaser

Table 3 presents the performance of our model targeting digital phaser under seen control parameter settings. Results show a consistent trend of increased ESR with higher feedback. As shown in the table, our model significantly outperforms the baseline model.

< **12** >

Table 3: *ESR(%) results for model trained on the digital phaser (Studio One Phaser[2]) under seen control parameter conditions. Training separate models for each control parameter configuration obtained results for the baseline model[4].*

LFO freq.(Hz)	Feedback(%)	ESR(%)	
		CONMOD	Baseline[4]
0.23	0	0.30	18.3
	25	0.21	19.6
	50	0.86	22.6
	75	4.34	32.2
0.73	0	0.08	19.1
	25	0.09	21.9
	50	0.21	23.7
	75	2.20	34.3
1.13	0	0.07	38.6
	25	0.08	22.4
	50	0.17	22.8
	75	1.96	49.8
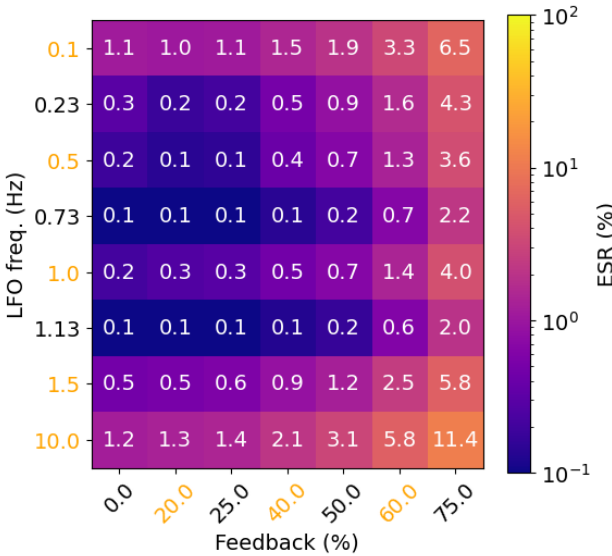
Figure 3: *ESR(%) results for model trained on the digital phaser under seen and unseen LFO frequencies and Feedback parameters. Seen cases are labeled black, and orange for unseen cases.*

Figure 3 shows the model inference score for both seen and unseen LFO frequencies and feedback during the training. Evaluations with unseen parameter configurations have been included to assess the model's controllability. We found that the model demonstrates robust inference capabilities for unseen LFO frequencies, even outside the range of the training data. This indicates that the LSTM in our model is capable of predicting the spectral behavior for the varying LFO frequencies. We also found that our model shows consistent performance for unseen feedback conditions. It is evident that the FiLM layer in our model demonstrates strong performance in facilitating continuous control inferences.

5.1.2. Analog Phaser

Figure 4 outlines the ESR results for the analog phaser model under seen control parameter settings. Similar to digital phaser, increased feedback leads to reduced accuracy. Notably, the baseline model failed to converge. We attribute this to the baseline structure learning model parameters assuming that the phaser is comprised of cascaded first-order all-pass filters.
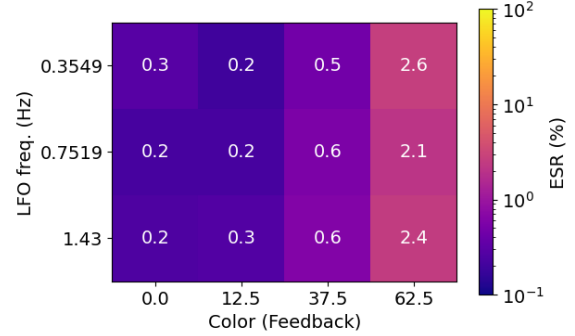

Figure 4: *ESR(%) results for model trained on the analog phaser under seen LFO frequencies and Color parameters.*

Figure 5 presents the model performance on unseen control parameter settings. The model shows effectiveness in predicting unseen feedback cases. However, it encounters challenges when inferring unseen LFO frequencies, especially those significantly higher than those observed within the seen frequency range. This difficulty may arise from complexities in modelling the nonlinear characteristics of analog circuits.
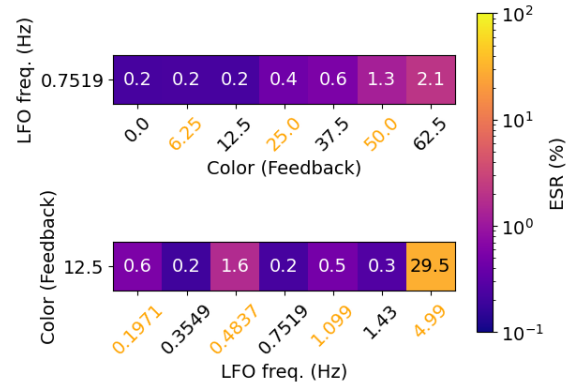

Figure 5: *ESR(%) results for model trained on the analog phaser under seen and unseen LFO frequencies and Color parameters. Seen cases are labeled black, and orange for unseen cases. Color is controlled over a fixed LFO frequency (Top). LFO frequency is controlled over a fixed Color rate (Bottom).*

5.1.3. Digital Flanger

The flanger effect demonstrates higher spectral complexity compared to the phaser, introducing greater number of notches in the spectrogram of transfer function[17]. As a result, the flanger exhibits an overall higher ESR than the phaser.
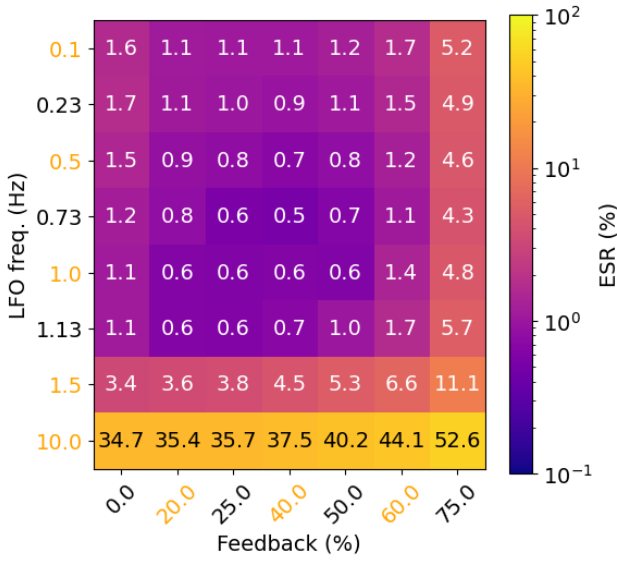
< **13** >

Figure 6: *ESR(%) results for model trained on digital flanger for under seen and unseen LFO frequencies and Feedback parameters. Seen cases are labeled black, and orange for unseen cases.*

Figure 6 displays the model's inference capability across both seen and unseen control parameters. The model exhibits strong performance for both seen and unseen feedback values. However, the model encounters difficulties in precisely predicting very high LFO frequencies, such as 10Hz. We anticipate that implementing adversarial loss [26] could mitigate these issues for the extreme control parameter settings, and leave this as a consideration for future work.

5.1.4. Analog Flanger

Figure 7 presents the ESR outcomes for the analog flanger model. Similar to previous experiments, the ESR decreases as feedback increases. Although the ESR is higher compared to the phaser scenario, it maintains an acceptable value overall.

Figure 8 illustrates the model's inference performance on unseen LFO frequency and feedback parameters. We observed that the model demonstrates robust interpolation for unseen feedback. Analogously, the model exhibited limitations in accurately inferring higher LFO frequencies.

5.2. Robustness on long-term audio inference

Robust long-term audio generation is essential for effect pedals, considering their use cases. Therefore, we analyzed our model's performance in long-term generation, with the results illustrated in Figure 10. In the case of the analog flanger, a minor increase in ESR was observed as the sequence length extended. However, in general, our model demonstrated the ability to generate stable sequences of up to 60 seconds. Due to the periodic nature of the LFO, the LSTM was able to perform inference effectively even as the sequence length increased.
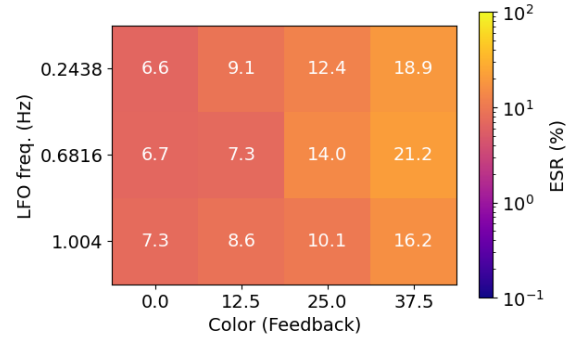

Figure 7: *ESR(%) results for model trained on the analog flanger under seen LFO frequencies and Color parameters.*

5.3. Effect of training the model on multiple LFO frequencies

To investigate the effects of training the model with multiple LFO frequencies, we compared the unseen LFO frequency inference capabilities of model trained with single LFO frequency and multiple LFO frequencies. As shown in Figure 11, the spectrogram of the audio output from the model trained with a single LFO frequency (left) exhibits an inaccurate rightward-shifted notch. In contrast, our proposed architecture leveraging multiple LFO frequencies (center) displays a notch shape identical to the target audio (right). This outcome aligns with expectations, due to the restricted diversity of LFO phases encountered when the model is trained using a single oscillator.
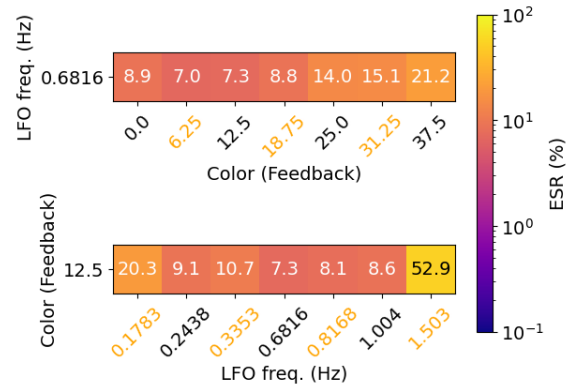

Figure 8: *ESR(%) results for model trained on the analog flanger under varying LFO frequencies and Color parameters. Seen cases are labeled black, and orange for unseen cases. Color is controlled over a fixed LFO frequency (Top). LFO frequency is controlled over a fixed Color rate (Bottom).*

5.4. Training on multiple phaser effects

To investigate the capability of our black-box modelling approach, we additionally trained our model concurrently on two distinct digital phasers: Studio One Phaser[2] and the Pedalboard Phaser[3]. The detailed control parameter settings for the training are described

---

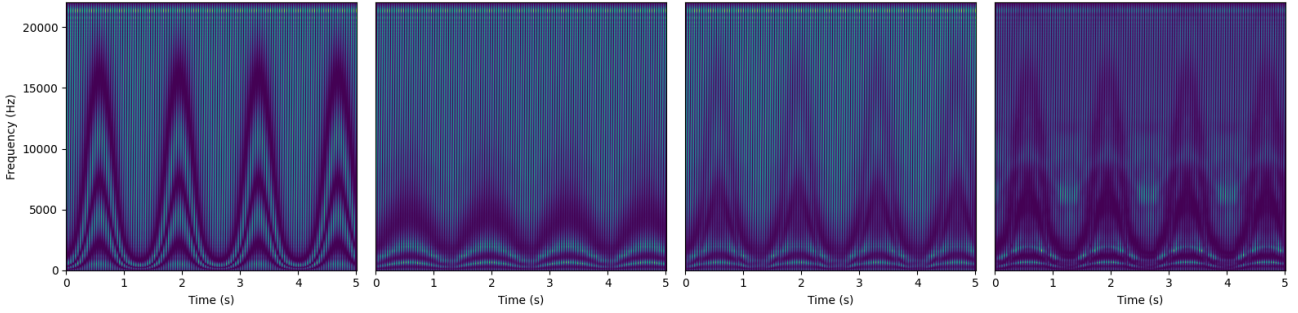[3]https://github.com/spotify/pedalboard

< **14** >

Figure 9: *Analysis of the multiple-phaser effects trained results, depicted in spectrograms. For better visualization, we used chirp signal as input. The model is simultaneously trained with two digital phaser effects: Studio One Phaser[2] and Pedalboard phaser. Each spectrogram illustrates the model's output with the Pedalboard phaser embedding(leftmost), the model's output with the Studio One Phaser[2] embedding(left), a naive linear sum of the outputs from each embedding(right), and model's output with the interpolated embedding(rightmost).*
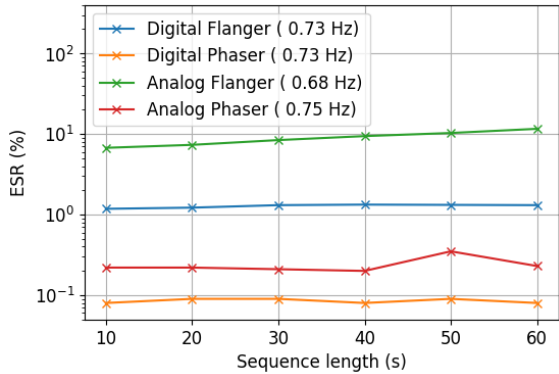

Figure 10: *ESR(%) for various input sequence length.*

Figure 11: *Analysis on the effect of training the model on multiple LFO frequencies. Each spectrogram illustrates the output of single LFO frequency model(left), multiple LFO frequency model(center), and target audio(right).*

in Table 4. The 2-dimensional embedding vector c_{emb} was concatenated with c_{fb} to be an input condition of the FiLM block. After training the model, we derived a unique embedding vector interpolated between the two vectors each representing individual effects and analyzed the model's output upon this condition. As illustrated in Figure 9, the model demonstrates consistently robust performance when conditioned with the learned embeddings for both phaser effects (leftmost and left). Notably, when the interpolated embedding is employed as the model condition (rightmost), the output exhibits a distinctive characteristic that lies between the two effects. For comparative analysis, we also present a result of a naive approach: a spectrogram of linearly summed audio from the two phaser effects (right).

6. CONCLUSION

In this study, we presented a black-box model for LFO-driven effects targeting phaser and flanger. Our model offers controllability over LFO frequency and feedback parameters, by employing a combination of LSTM, MLP, and FiLM blocks, along with a unique training procedure optimizing multiple LFO frequencies. Comparative analysis with a baseline model revealed our model's superiority in both seen and unseen control parameters when modelling the phaser effect. Additionally, we found that our model can effectively generate audible flanger effects with reasonable con-
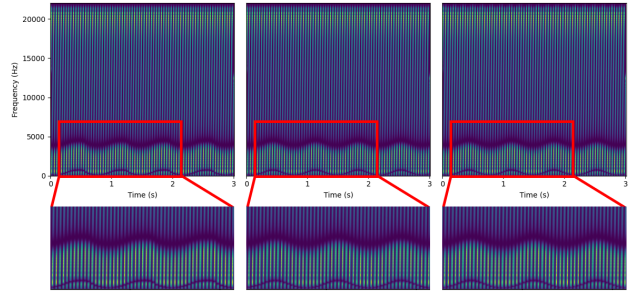
Table 4: *Control parameter experiment settings for training multiple phaser effects.*

Studio One Digital Phaser (Phaser[2])	
Stages	6
Speed (Hz)	0.37, 1.13
Center	1300
Range	80%
Feedback (%)	0, 25, 50, 75
Stereo Spread	2.0%
Mix	70.0%
Pedalboard Phaser	
Rate (Hz)	0.37, 1.13
Centre Frequency (Hz)	1300
Depth	0.5
Feedback	0.0, 0.25, 0.5, 0.75
Mix	0.5

trollability. Also, our model is capable of generating long sequences consistently with negligible amount of decrease in accuracy. Finally, we demonstrated our model's capability to train on two distinct phaser effects concurrently, resulting in creative mixed-up outputs between the effects.

However, our model faced challenges in inferring LFO frequen-

< **15** >

cies that were outside the training data range, particularly evident in analog effect modelling. The performance was relatively lower for the flanger effect, due to its higher spectral complexity. Addressing these challenges may involve implementing methods capable of modelling complex distributions, such as adversarial loss. We aim to explore these solutions in future work.

7. REFERENCES

[1] Tara Vanhatalo, Pierrick Legrand, Myriam Desainte-Catherine, Pierre Hanna, Antoine Brusco, Guillaume Pille, and Yann Bayle, "A review of neural network-based emulation of guitar amplifiers," *Applied Sciences*, vol. 12, no. 12, 2022.

[2] Riccardo Simionato and Stefano Fasciani, "Fully conditioned and low-latency black-box modeling of analog compression," in *Proceedings of the International Conference on Digital Audio Effects*. DAFx Board, 2023.

[3] Marco Comunità, Christian J Steinmetz, Huy Phan, and Joshua D Reiss, "Modelling black-box audio effects with time-varying feature modulation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[4] Alistair Carson, Cassia Valentini-Botinhao, Simon King, and Stefan Bilbao, "Differentiable grey-box modelling of phaser effects using frame-based spectral processing," in *International Conference on Digital Audio Effects (DAFx)*, 2023.

[5] Jong Wook Kim, Rachel Bittner, Aparna Kumar, and Juan Pablo Bello, "Neural music synthesis for flexible timbre control," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 176–180.

[6] Christian J Steinmetz and Joshua D Reiss, "Steerable discovery of neural audio effects," in *5th Workshop on Creativity and Design at NeurIPS*, 2021.

[7] Felix Eichas, Marco Fink, Martin Holters, and Udo Zölzer, "Physical modeling of the mxr phase 90 guitar effect pedal.," in *DAFx*, 2014, pp. 153–158.

[8] Jaromír Mačák, "Simulation of analog flanger effect using bbd circuit," in *Proceedings of the 19th International Conference on Digital Audio Effects (DAFx-16), Brno, Czech Republic*, 2016, pp. 5–9.

[9] Thomas Hélie, "Volterra series and state transformation for real-time simulations of audio circuits including saturations: Application to the moog ladder filter," *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 4, pp. 747–759, 2009.

[10] Felix Eichas and Udo Zölzer, "Black-box modeling of distortion circuits with block-oriented models," in *Proceedings of the International Conference on Digital Audio Effects (DAFx), Brno, Czech Republic*, 2016, pp. 5–9.

[11] Marco A Martínez Ramírez, Emmanouil Benetos, and Joshua D Reiss, "A general-purpose deep learning approach to model time-varying audio effects," in *International Conference on Digital Audio Effects (DAFx)*, 2019.

[12] Roope Kiiski, Fabián Esqueda, and Vesa Välimäki, "Time-variant gray-box modeling of a phaser pedal," in *Proc. Int. Conf. Digital Audio Effects (DAFx-16), Brno, Czech Republic*, 2016, pp. 31–38.

[13] Joseph T Colonel, Marco Comunità, and Joshua Reiss, "Reverse engineering memoryless distortion effects with differentiable waveshapers," in *Audio Engineering Society Convention 153*. Audio Engineering Society, 2022.

[14] M Martinez Ramirez, Joshua Reiss, et al., "End-to-end equalization with convolutional neural networks," 2018.

[15] Marco A Martínez Ramírez, Emmanouil Benetos, and Joshua D Reiss, "Deep learning for black-box modeling of audio effects," *Applied Sciences*, vol. 10, no. 2, pp. 638, 2020.

[16] Christian J Steinmetz and Joshua D Reiss, "Efficient neural networks for real-time modeling of analog dynamic range compression," in *152nd Convention of the Audio Engineering Society (AES)*, 2022.

[17] Alec Wright and Vesa Välimäki, "Neural modelling of periodically modulated time varying effects," in *International Conference on Digital Audio Effects*. DAFx, 2020, pp. 281–288.

[18] Alec Wright and Vesa Välimäki, "Neural modeling of phaser and flanging effects," *Journal of the Audio Engineering Society*, vol. 69, pp. 517–529, 07 2021.

[19] Christopher Mitcheltree, Christian J. Steinmetz, Marco Comunità, and Joshua D. Reiss, "Modulation extraction for lfo-driven audio effects," in *International Conference on Digital Audio Effects (DAFx)*, 2023.

[20] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, 2018, vol. 32.

[21] Michael L Beigel, "A digital phase shifter for musical applications, using the bell labs (alles-fischer) digital filter module," *Journal of the Audio Engineering Society*, vol. 27, no. 9, pp. 673–676, 1979.

[22] "Studio one manual," https://s1manual.presonu s.com/en/StudioOneReferenceManual.htm, Accessed: 2024-03-23.

[23] "Flamma fc15 analog flanger pedal manual," https://fl ammainnovation.com/file/pdf?url=group1/M 00/00/07/wKgA0GNFRIeAX4hnAAsXY4lMquE867. pdf&fileName=FC15_Manual_EN_V01_2022.10. 11.pdf, Accessed: 2024-03-23.

[24] Ben Hayes, Charalampos Saitis, and György Fazekas, "Sinusoidal frequency estimation by gradient descent," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[25] Aaron Van Den Oord, Oriol Vinyals, et al., "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[26] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17022–17033, 2020.

< **16** >