IMPROVING UNSUPERVISED CLEAN-TO-RENDERED GUITAR TONE TRANSFORMATION USING GANS AND INTEGRATED UNALIGNED CLEAN DATA

Yu-Hua Chen

Graduate Institute of Networking and Multimedia
National Taiwan University
Taipei, Taiwan
f08946011@ntu.edu.tw

Woosung Choi, Wei-Hsiang Liao*, Marco A. Martínez-Ramírez*, Kin Wai Cheuk*, Yuki Mitsufuji*,†*

* Sony AI, † Sony Group Corporation
Tokyo, Japan
woosung.choi@sony.com

Jyh-Shing Roger Jang

Graduate Institute of Networking and Multimedia
National Taiwan University
Taipei, Taiwan
jang@mirlab.org

Yi-Hsuan Yang

Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan
yhyangtw@ntu.edu.tw

ABSTRACT

Recent years have seen increasing interest in applying deep learning methods to the modeling of guitar amplifiers or effect pedals. Existing methods are mainly based on the supervised approach, requiring temporally-aligned data pairs of unprocessed and rendered audio. However, this approach does not scale well, due to the complicated process involved in creating the data pairs. A very recent work done by Wright *et al.* has explored the potential of leveraging unpaired data for training, using a generative adversarial network (GAN)-based framework. This paper extends their work by using more advanced discriminators in the GAN, and using more unpaired data for training. Specifically, drawing inspiration from recent advancements in neural vocoders, we employ in our GAN-based model for guitar amplifier modeling two sets of discriminators, one based on multi-scale discriminator (MSD) and the other multi-period discriminator (MPD). Moreover, we experiment with adding unprocessed audio signals that do not have the corresponding rendered audio of a target tone to the training data, to see how much the GAN model benefits from the unpaired data. Our experiments show that the proposed two extensions contribute to the modeling of both low-gain and high-gain guitar amplifiers.

1. INTRODUCTION

Amplifier modeling involves developing algorithms to emulate the behavior of real amplifiers. The amplifiers typically discussed in the literature are vacuum tube amplifiers. This task can also be considered a virtual analog (VA) modeling problem. Recent studies have demonstrated the potential to apply neural networks to VA modeling tasks using supervised learning. Various network architectures have been proposed in the literature, such as convolution-based and recurrent-based networks [1, 2, 3, 4, 5].

Training in a supervised setting has already yielded promising results for VA modeling and guitar amplifier modeling. Many commercial applications have adopted this approach, using mainly the minimization of the error-to-signal ratio (ESR) as the training objective [4, 6]. However, the supervised approach does not scale well, for it requires paired data to model the transformation process from a clean audio input to a rendered audio output for a target tone. Each data pair has to be temporally aligned and be about the same content of guitar performance. In many cases, however, the tone or timbre of a *rendered* audio signal often lacks the *unprocessed*, or direct-input (DI), audio counterpart, making supervised methods impractical. While it might be possible to create such data pairs by inverting the clean audio directly from a rendered audio by means of a guitar effect removal model, the development of such models is still an ongoing area of research [7].

In other audio synthesis tasks such as neural vocoding [8, 9] and voice conversion [10, 11], many advanced generative adversarial network (GAN) [12] models have been developed to generate realistic waveforms. For example, MelGAN [8] proposed a multi-scale discriminator (MSD) for distinguishing between real audio and generated audio. HiFi-GAN [13] proposed a multi-period discriminator (MPD) that collaborates with the MSD. Compared to models that minimize directly the reconstruction loss, GAN models employ such discriminators to learn customized loss functions in a data-driven fashion, usually leading to models that generate audio with finer details and better perceptual quality empirically.

We conjecture that a GAN-based approach can similarly offer two advantages for guitar amplifier modeling.

Adversarial losses Adversarial losses offer a way to learn complicated, high-dimensional probability distributions from diverse and high-quality training data samples without explicitly modeling the underlying probability density function. Specifically, GANs implicitly learn the data distribution using a self-learned loss function that is dynamically-adjusted as the training process unfolds.

Unsupervised training We can use any available unpaired clean data as input to the generator during the training process, with the target being the designated amplifier-rendered data, thus potentially improving the generalizability of the model while reducing the burden of collecting paired data.

To the best of our knowledge, the work of Wright *et al.* [14] represents the first and the only existing work that adopts GANs for guitar amplifier modeling. Viewing amplifier modeling as a style transfer problem, they showed that a GAN-based model using the

MSD proposed in MelGAN [8] as the discriminator can learn the amplifier modeling process without using reconstruction loss functions such as the ESR. Moreover, they conducted experiments involving mismatched guitar timbre conversion between two timbres produced from distinct guitars. These experiments demonstrate the potential of adapting the unsupervised approach for guitar amplifier modeling.

Being inspired by the work of Wright *et al.* [14], we set forth to further extend this GAN-based approach by presenting the following two extensions. First, while they position their work in the context of audio *style transfer* by employ MelGAN[8] and several choices of spectral discriminator, we further highlight the potential of integrating more advanced discriminators as proposed in neural vocoder research. For example, it is well known that HiFi-GAN [13] empirically generates audio waveforms with higher quality than MelGAN [8]. Research on neural vocoders is relevant, because both vocoders and guitar amplifier modeling aim to produce high-quality audio waveforms given some input conditions. Consequently, our first extension replaces the MSD discriminator used in [14] by a combination of MSD and MPD discriminators, to study whether advanced discriminators can similarly contribute to better result for guitar amplifier modeling as the case seen in neural vocoding.

Our second contribution investigates more deeply the benefits of a GAN-based model in utilizing unpaired data. Specifically, we note that during the training process, Wright *et al.* [14] only used the unprocessed audio that *do* have the corresponding rendered audio of the target tone as the input to the generator. However, as the GAN training does not require paired data, it is actually possible to utilize unprocessed audio that *do not* have the rendered audio counterpart of the target tone as the generator's input. We study such a case in our work, using input audio signals that do not align with the target output audio signals in training our model.

We conduct experiments on two public-domain guitar datasets, the *EGDB* dataset [15] that have both low-gain and high-gain tones, and the *EGFxset* dataset [16] for an extremely high-gain tone. Experimental results show that the proposed extensions contribute positively to the modeling result, especially for the extremely high-gain case. We provide audio samples online.[1]

The paper is structured as follows: Section 2 reviews neural amplifier modeling methods and GANs. Section 3 presents our proposed method. Section 4 describes the dataset and experimental setup; Section 5 reports the objective evaluation results. Section 6 discusses the results further. Finally, Section 7 concludes the paper with some ideas of future work.

2. RELATED WORKS

2.1. Neural Amplifier modeling

Thanks to advancements in deep learning, neural networks have been utilized in several studies on amplifier modeling [4, 6, 17, 18, 19, 20]. The neural network approach shares similarities with traditional black-box methods. For example, a convolutional layer can be conceptualized as a Wiener model [21]. Existing neural network models for amplifier modeling are usually adapted from neural network models that are initially proposed for speech-related tasks. While speech signals commonly operate at a 16 kHz sampling rate, overdrive or distortion sound characteristics frequently manifest in the higher frequency range, requiring sampling rates of 44.1 kHz or 48 kHz. As such, the adaptations may result in increased complexity and model size, which can be unfavorable given the requirements on real-time efficiency and low latency of VA modeling.

2.2. Generative Adversarial Networks

A GAN [12] is a generative model contains two components: a generator G and a discriminator D. The discriminator D is essentially a classifier and it aims to output a value close to 1 for samples from "real" data distribution $x \sim p_d$, and a value close to 0 for "fake" samples $G(z)$ generated by the generator G, whose input z is sampled from a prior distribution p_z. On the other hand, the generator seeks to deceive the discriminator by generating samples that are indistinguishable from real ones. The two-player minimax game with the value function $V(G, D)$ is defined as follows, updating G and D iteratively as the training unfolds,

$$\begin{aligned} \min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\mathrm{d}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \\ \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))] \end{aligned} \tag{1}$$

For VA modeling, it is the generator G that performs the clean-to-rendered transformation during both the training and inference stages. The discriminator D only functions during the training stage, guiding how the generator is optimized. Therefore, it is possible to use a computationally heavy discriminator to train a light generator, for better run-time efficiency of the generator.

2.3. Backbone Model for Generator

Existing approaches to neural VA modeling can be categorized into two main types: convolutional (CNN) and recurrent neural networks (RNN). From a digital signal processing viewpoint, CNN networks can be viewed as finite impulse response (FIR) filters. CNN-based models have demonstrated superior performance in modeling various devices. For example, Wright *et al.* [1] applied a WaveNet model [22] to model the Blackstar HT-5 Metal and the Mesa Boogie 5:50 Plus amplifiers. Damskägg *et al.* [2] utilized a WaveNet model with conditioning control on the gain parameter to emulate the Fender Bassman 56F-A vacuum-tube amplifier. In addition to amplifier modeling tasks, Steinmetz *et al.* [3] trained a conditional temporal convolutional network on compressor, analog delay, guitar amplifier, and reverberation effects.

On the other hand, RNN-based approaches often rely on long-short term memory (LSTM) or gated recurrent units (GRU). For instance, Wright *et al.* [4] showed promising results using a recurrent-based model to model a high-gain channel Blackstar HT-1 vacuum tube amplifier and an Electro-Harmonix Big Muff Pi distortion/fuzz pedal. Juvela *et al.* [5] extended their work further by concatenating control parameters with a range of [0,1] as additional input channels to their LSTM network.

2.4. Discriminators for GANs training

To apply adversarial losses within the GAN framework, a discriminator is needed to distinguish between real data and generated output. Several discriminators have been proposed for audio generative tasks such as neural vocoder [8, 13], voice conversion [10, 11], and neural codec [23]. We categorize these discriminators into two types: spectral-based and waveform-based discriminators.

For spectral-based discriminators, Défossez *et al.* [23] proposed a a multi-scale STFT-based discriminator. They computed

[1] `https://ampDaFX24.notionlinker.com`

	GuitarSet	EGDB	GUITAR-FX-DIST	EGFxSet
Clean	3h	2h	0.57h	∼1h
Rendered	N/A	10h	∼111h	11.5h

Table 1: The total duration (in hours, or 'h') of the clean, unprocessed audio and the rendered audio (with effects applied) of four existing public paired datasets, GuitarSet [24], EGDB [15], GUITAR-FX-DIST [25], and EGFxSet [16].

and summed the short-time Fourier transform (STFT) losses with different parameters (i.e., FFT size, window size, and hop length). These techniques compel the generator to not only focus on generating the waveform itself but also to generate reasonable results in the spectral domain. On the other hand in MelGAN [8], a multi-scale discriminator (MSD) was introduced, operating on different audio scales (i.e., sample rates) in waveform domain. Each scale's audio is processed by a 1-D convolution-based module to obtain an output. The outputs from each scale are then used to calculate adversarial losses for training the discriminator. In Hifi-GAN [13], a multi-period discriminator (MPD) was proposed to capture both regular and prime number distances between sample points, resulting in improved speech synthesis quality. While the previous GAN-based VA modeling model [14] uses MSD alone, we use MSD as well as MPD in our work.

2.5. Clean Audio from Existing Datasets

We refer to a dataset as a *paired dataset* when it contains the clean audio signal counterpart for each amplifier- or pedal-rendered audio signal. Referring to existing public-domain paired datasets, we show a comparison of clean audio and rendered audio duration in Table 1. As the table shows, the duration of amplifier or pedal rendered data in each dataset is much longer than the aligned clean audio in overall duration. We see that clean audio is relatively scarcer than rendered audio.

We categorize the clean audio into two types during training. First is *target-aligned* clean audio, where a target audio can always find an aligned clean audio with the same musical content. Second is *target-unaligned* clean audio, where the musical content in this type of clean audio does not exist in target amplifier-rendered audio. Please note that the objective evaluation metrics such as ESR and Mel-spectrum loss (cf. Section 4.2) still requires data from a target-aligned setting.

As mentioned in Section 1, even though the clean data and rendered data from the *same* dataset is usually fully aligned (i.e., they are target-aligned), we can take advantage of the GAN-based approach and further use clean data and rendered data from *different* datasets and employ such target-unaligned data in our unsupervised training. The prior work of Wright *et al.* [14] did not exploit such a potential, as they used clean audio from [26] and created rendered target audio from three different plugins, essentially creating target-aligned data. Unlike their work, we study the use of target-unaligned data in our experiments.

3. METHODS

We consider guitar amplifier modeling as a generative task that aims to generate high-fidelity audio waveforms. Given an input audio of T samples, $\mathbf{x} \in \mathbb{R}^{1\times T}$, we adopt the "black-box" approach and train a neural network-based generator G that carries out the amplifier modeling process and generates $\hat{\mathbf{x}} = G(\mathbf{x})$. We illustrate our training framework in Figure 1.

Model	channels	kernel sizes	stride	groups	padding
conv1d	(1, 128)	15	1	1	0
conv1d	(128, 128)	41	2	4	20
conv1d	(128, 256)	41	2	16	20
conv1d	(256, 512)	41	4	16	20
conv1d	(512, 1024)	41	4	16	20
conv1d	(1024, 1024)	41	1	16	20
conv1d	(1024, 1024)	5	1	2	0

Table 2: Parameter settings of the convolutional layers of the implemented MSD sub-discriminators.

Model	channels	kernel sizes	stride	groups	padding
conv2d	(1, 32)	(5, 1)	(3,1)	1	2
conv2d	(32,128)	(5, 1)	(3,1)	1	2
conv2d	(128, 512)	(5, 1)	(3,1)	1	2
conv2d	(512, 1024)	(5, 1)	(3,1)	1	2
conv2d	(1024, 1024)	(5, 1)	(1,1)	1	2
conv2d	(1024, 1)	(1,1)	(3,1)	1	2

Table 3: Parameter settings of the convolutional layers of the implemented MPD sub-discriminator.

3.1. Generator

We employ the same causal feed-forward WaveNet model architecture as Wright *et al.* [14] for our generator. It consists of two stacks of nine dilated convolution layers. The dilation is one at the first stack and is increased by a factor of two after each stack to get a larger receptive field. We set a kernel size as 3 to get a growth receptive field from small area to larger area. Each convolution layer is equipped with a weight normalization. We use the same gated activation function as the original WaveNet model [22].

3.2. Discriminator

Our discriminator consists of both MSD- and MPD-based ones. The MSD consists of three sub-discriminators originally [13]. However, we remove the last sub-discriminator that processes the audio after two downsampling layers (i.e., the one that operates at the lowest temporal resolution), as this gives better results empirically in our pilot study. The input flow for MSD is therefore: raw audio, the first sub-discriminator, $\times 4$ average-pooled audio, and finally the second sub-discriminator. We set the parameters as shown in Table 2. Following the setting of Hifi-GAN [13], spectral normalization is applied for the first sub-discriminator, while weight normalization is applied for the second one.

The MPD comprises a collection of mixture sub-discriminators. Unlike MSD, it only accepts equally-spaced sample points of the input audios. With audio length T and period P, input for each sub-discriminator will be reshaped from the audio length T to $(T/P, P)$ (i.e., from 1D to 2D). Following the setting of Hifi-GAN [13], we employ multiple sub-discriminators, each operating with a period p in $P = [2, 3, 5, 7, 11]$. We set the parameters of the convolutional layers as shown in Table 3. Each sub-discriminator is a stack of convolutional layers, with weight normalization applied for every convolutional layers. This setup allows us to obtain a set of discriminative outputs for each period sub-discriminator, providing different perspectives based on different period settings.

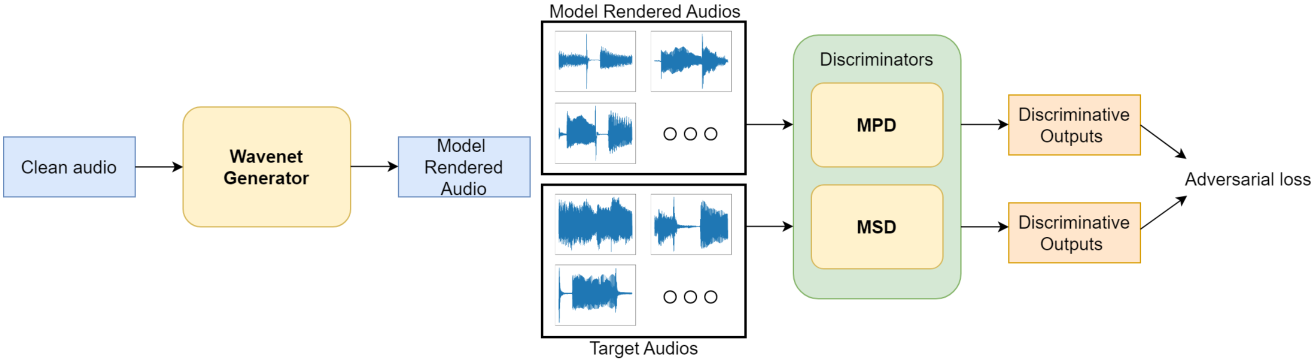

Figure 1: Diagram of the proposed GAN-based model for VA modeling, using clean audio that may not be matched and aligned with the target audio segment, and using two types of discriminators: MSD and MPD [13].

There are also other voice synthesis works under GANs framework collaborating their discriminators with spectral-based discriminator [27, 28]. Either a hierarchical [28] or multi-resolution [27] approach aims to capture the information from different perspectives of sample rates or window sizes of short-time Fourier transform. We opt to follow several settings from [13] not solely due to its success in speech synthesis. In our preliminary experiments, we found that applying MPD as a module of the discriminator can help model high frequency information in distortion or overdrive audio. Furthermore, the combination of MSD+MPD has been not only applied in neural vocoder tasks but also in neural codec tasks [9]. In summary, the main goal of these audio-related tasks is to generate high-fidelity sound, and utilizing multi-type discriminators can improve audio fidelity under GAN training.

3.3. GAN Loss

We choose to apply a Hinge GAN loss function [29] in our GAN training, due to its promising result in prior work [9, 14]. The loss equation is defined as follows:

$$\mathcal{L}(D;G) = \mathbb{E}_y[\max(0, 1 - D(y))] + \mathbb{E}_x[\max(0, 1 + D(G(x)))] \tag{2}$$

$$\mathcal{L}(G;D) = \mathbb{E}_x[-D(G(x))] \tag{3}$$

During training, the discriminator is trained to classify labeled data y as 1, and the samples generated from Generator $G(x)$ as 0. The generator is trained to deceive the discriminator into recognizing $G(x)$ as real data, aiming for a classification close to 1. Other auxiliary losses such as *mel-spectrogram loss* and *feature matching loss* were used in [9, 30, 8]. The mel-spectrogram loss measures the L1 distance between the generated audio's mel-spectrogram and that of labeled data. The feature matching loss computes the L1 distance in intermediate features from the discriminators between the generated audio and a labeled data. Although these loss functions can improve the training efficiency, stability of the generator, and the quality of the generated audio, they require a paired data setting during the training process. Given an unpaired data setting, our model only utilizes an adversarial loss during the training process for both the generator and discriminator.

4. EXPERIMENTAL SETUP

4.1. Dataset

We select two electric guitar datasets for our experiments: EGDB [15] and EGFxset [16]. For EGDB, the duration of a single tone is approximately 2 hours. We choose a subset consisting of **Marshall JCM2000**, **Fender Twin Reverb**, and **Mesa Boogie Mark**. For EGFxset, we select the **BD-2** dataset as the target tone for our experiments. As the gain value has been set to its maximum value for BD-2, this dataset contains highly distorted sound. As the BD-2 tone has fairly high gain, we may consider the three tones from EGDB as relatively low-gain tone compared to BD-2. We note that EGDB comprises musical phrases or licks, while EGFxset contains recordings of individual notes, each at different pitches and from various pickups.

In our preliminary experiments, we found great differences in amplitude between the two datasets, possibly because they were collected under different device settings (e.g., guitar or audio interface) and recording environments. We found that GAN-based models is highly sensitive to differences in amplitude. To address this, we normalize both datasets using pyloudnorm [31]. Specifically, we normalize the peak of each audio to -1 dB, and then normalized each audio to -12 dB LUFS. Without such a normalization, the training would be extremely unstable, leading to failures during the early stages of the GAN training process.

We divide the dataset into training, validation, and test sets using an 80/10/10 ratio. For training, the input clean data and the output target tone data are randomly arranged in each batch for an unsupervised setting. To evaluate the model performance, as the clean data and rendered data are aligned between the two datasets, validation and testing are conducted under a paired setting to calculate all metrics.

4.2. Metrics

We consider the following three metrics for objective evaluation.

Error-to-signal ratio (ESR) is a metric commonly employed for training and evaluating an amplifier modeling model. For N sample points, a pre-emphasis filter is applied to both the generated signal $\widehat{\mathbf{y}}_p$ and a target signals $\mathbf{y}_p$ before computing the ESR.

$$\mathrm{ESR} = \frac{\sum_{n=0}^{N-1} |\mathbf{y}_p[n] - \widehat{\mathbf{y}}_p[n]|^2}{\sum_{n=0}^{N-1} |\mathbf{y}_p[n]|^2}.$$

The denominator of target signal itself is used to prevent the high energy segments from dominating the result.

Mel-spectrum loss ($L1_{mel}$) measures the difference between the ground-truth and predicted audio in the spectral domain. Denoting $\phi(\cdot)$ as the function that converts an audio waveform into a Mel spectrogram, this loss can be calculated as follows,

$$L1_{mel} = \mathbb{E}_{(\mathbf{x})} \left[\|\phi(\mathbf{x}) - \phi(\hat{\mathbf{x}})\|_1 \right] .$$

Since the characteristics of certain effects such as overdrive and distortion are more easily observed from the perspective of spectral domain, we conjecture that Mel-spectrum loss offers a more suitable measure than ESR for evaluating the performance of modeling overdrive and distortion.

Fréchet Audio Distance (FAD) [32] measures the Fréchet distance between the distribution of embedding from a set of reference audios and those from the generated audios. It was first proposed to evaluate a music enhancement task and was found to correlate well with human perception. The metric has been later applied to music generation tasks (e.g., [33, 34, 35]) to indicate if the generated audio is plausible. We accordingly adopt it here as well. To provide different insights than traditional alignment-based metrics, we report the FAD[2] score of all models with the VGGish model. Samples with a low FAD score are expected to be more plausible. We note that there might be better alternatives than the VGGish model for computing the FAD scores [36], but we leave that as a future work.

4.3. Implementation details

For detailed information on the generator and discriminator, please refer to Sections 3.1 and 3.2. It is important to note that we use the same structures and settings for all generator models, including the supervised method, for fair comparison.

During training, we split every audio into two-second segments. Each model is trained using a single RTX 3090 GPU. The supervised and MelGAN discriminator models are trained with the settings described in their original papers.

For our discriminators, MSD and MSD+MPD, we employ the AdamW [37] optimizer with an initial generator learning rate of 5e–5 and discriminator learning rate of 1e–5, along with a weight decay of 0.01. We set the generator learning rate empirically to a higher value to prevent an imbalance in the training process between the generator and discriminator, as our generator model sizes are much smaller than those of the proposed discriminators.

4.4. Evaluation settings

We conduct two experimental scenarios to evaluate the introduced three metrics mentioned in Section 4.2. First, we compare our method with the supervised approach of Damskägg *et al.* [6] and the unsupervised GAN-based approach proposed by Wright *et al.* [14]. Second, to validate the advantage of unsupervised learning, we combine the clean audio from EGDB and EGFxset as the input to our generator (marked as "both" in Table 4), no matter whether the target tone is from EGDB or EGFxset. The generator for all the aforementioned models used the same architecture; the difference in each model setting lies in the training method and the loss functions applied (e.g., supervised or adversarial).

[2] To compute FAD, we use an open-source implementation: `https://github.com/gudgud96/frechet-audio-distance`

Target tone	Input	Model	$L1_{mel}$↓	ESR↓	FAD↓
BD-2 (EGFxset)	EGFxset	Supervised [6]	4.041	0.106	5.256
	EGFxset	MSD	1.874	0.164	1.900
	EGFxset	MSD+MPD	1.535	0.052	0.983
	both	MSD+MPD	**1.156**	**0.022**	**0.550**
Marshall (EGDB)	EGDB	Supervised [6]	2.342	**0.019**	1.657
	EGDB	MSD	2.660	0.229	1.410
	EGDB	MSD+MPD	**2.315**	0.028	**0.994**
	both	MSD+MPD	2.458	0.029	1.054
FTwin (EGDB)	EGDB	Supervised [6]	**1.953**	**0.014**	1.126
	EGDB	MSD	2.302	0.072	0.878
	EGDB	MSD+MPD	1.960	0.021	**0.346**
	both	MSD+MPD	2.267	0.020	0.434
Mesa (EGDB)	EGDB	Supervised [6]	1.705	**0.012**	1.923
	EGDB	MSD	2.158	0.137	1.674
	EGDB	MSD+MPD	**1.633**	0.014	1.748
	both	MSD+MPD	1.694	0.041	**1.400**

Table 4: Objective evaluation result of the supervised models [6] and GAN-based models (i.e., MSD alone and ours) for different target tones (i.e., an extremely high-gain tone from EGFxset and three tones from EGDB), using source signals from different datasets as the model input (i.e., using target-aligned audio, or using "both" target-aligned audio and target-unaligned audio [i.e., EGDB+EFGxset]). All the three objective metrics are the lower the better; best results highlighted in bold.

5. EXPERIMENTAL RESULT

5.1. Comparison with Baseline Methods

Table 4 shows the results of modeling different target tones. We consider firstly the case when the input audio are from the same dataset as the target audio, namely using EGFxset input when the target tone is BD-2, and using EGDB input when the target tone is the other three. We will consider the result for the case when using input from "both" EGFxset and EGDB in the next subsection.

As we address a challenging case with BD-2 as the target tone, which is characterized by a highly distorted tone, we opt for using the MelGAN discriminator as our baseline. This choice is based on its superior performance in handling heavy distortion settings, as demonstrated the experiments of Wright *et al.* [14], viewing the MelGAN discriminator used by them as a variant of MSD. Table 4 shows that the proposed GAN-based approach (i.e., MSD+MPD) consistently outperforms the existing GAN-based approach (i.e., MSD) across all the three objective metrics. Notably, when MPD and MSD were applied in GAN training, these discriminators, originally designed for capturing the diverse periodic patterns, helped the generator produce a more realistic waveform. Although our GAN-based model does not include any spectral-based discriminators, for $L1_{mel}$, our method shows a slight improvement on the low-gain tones from EGDB. In general, this result suggests that the combination of MSD and MPD leads to better VA modeling than MSD alone.

Table 4 also shows that, compared to the supervised baseline [6], the proposed GAN-based approach (MSD+MPD) does not lead to better results in ESR. However, for the challenging case of the high-gain tone BD-2 from the EGFxset, the proposed GAN-based approach outperforms the supervised baseline [6] greatly in all the three metrics, especially for $L1_{mel}$ and FAD. Informal listening to the generation result (examples available on the demo page) also shows that the proposed model performs perceptually

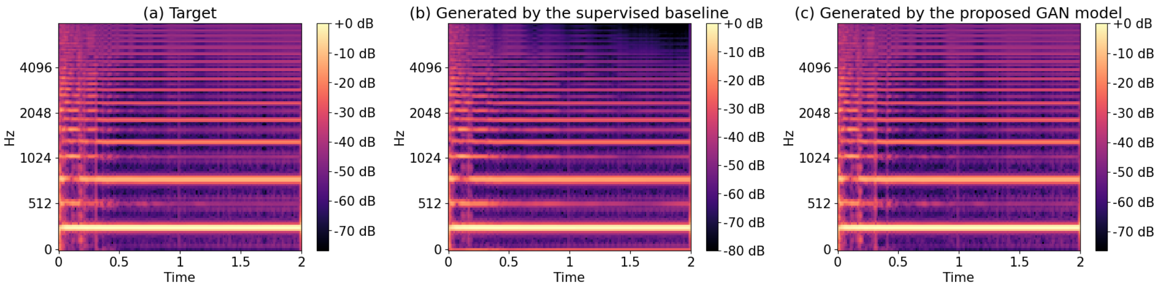

Figure 2: Mel-spectrogram of the target audio signal, along with the ones generated by the supervised baseline [6] and the proposed MSD+MPD GAN-based model given the corresponding clean audio signal. The target audio is sampled from the test set of EGFxset [16], with BD-2 being the target tone. We see missing high-frequency harmonics from the top-right corner of the middle Mel-spectrogram, the one generated by the supervised baseline.

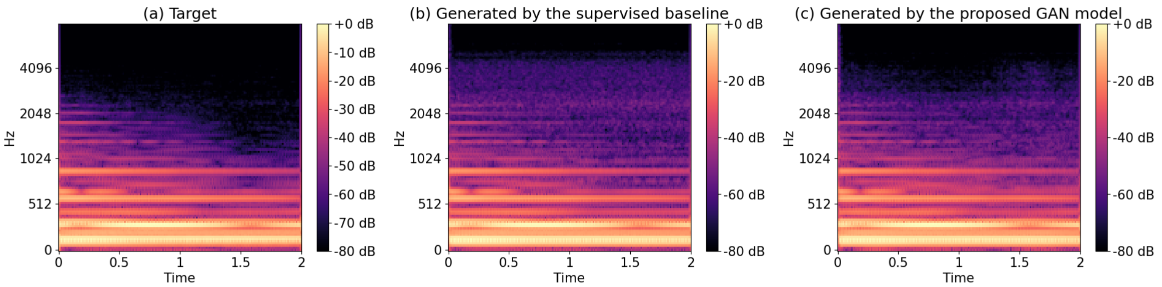

Figure 3: The mel-spectrogram between target audio , supervised approach and MSD+MPD sampled from the EGDB Fender test set. Both of supervised baseline and our MSD+MPD exhibit artifacts. These artifacts manifest as the generation of non-existent high-frequency information in the target mel-spectrogram.

better.[3] Using MSD alone also outperforms the supervised baseline in $L1_{mel}$ and FAD. Together, this suggests the advantage of the GAN-based loss for high-gain tones.

5.2. Clean Audio Combination

Next, we consider the case where we use the clean audio from both dataset as the input to our generator, no matter whether the target tone is from EGDB or EGFxset. Table 4 shows that applying this "both" input data setting can further boost the performance of all metrics using a paired dataset in a high-gain BD-2 target tone. In other words, the incorporation of EGDB clean tones contributes positively to the modeling of the target tone from EGFxset. This result provides more empirical evidences of the advantage of the GAN-based approach.

Interestingly, in contrast, we observe that incorporating clean data from EGFxset does not significantly contribute to modeling any of the target tones from EGDB. We conjecture that this is due to the differences in music content between the two datasets (e.g., licks versus single notes). Since EGFxset does not have note sequences and transitions between individual notes, adding inputs from EGDB could offer advantages. However, this is not the case when considering the reverse scenario.

[3] While we have not conducted a formal subjective evaluation of the implemented models, the listening test reported by the prior work of Wright *et al.* [14] has suggested that their GAN-based model (i.e., using MSD alone for the discriminator) outscores supervised models in human evaluation.

6. DISCUSSION

6.1. Benefit of the GAN-based Approach for VA Modeling

Due to the limits in the amount of available paired data and computation resource, we have only considered two datasets and four tones in total in our experiments. However, the experimental result has already revealed potentials of the GAN-based approach over the prevailing supervised approach for VA modeling. To illustrate this point further, we consider the tone that the supervised baseline [6] does not perform well, namely, the EGFxset BD-2, and conduct a case analysis.

Figure 2 plots the mel-spectrogram of a sampled target audio rendered with the BD-2 tone from EGFxset, along with the mel-spectrograms of the generation result of the supervised baseline [6] and the proposed model given the corresponding clean signal. We can see that many high-frequency harmonics are missing in the result of the supervised baseline. In contrast, they are effectively captured by the proposed model throughout the time axis. Similar observations can be found for other samples for the BD-2 tone.

From the viewpoint of digital signal processing (DSP), higher gain value implies more non-harmonic high-frequencies in the audio. While such non-harmonic high-frequencies may not be well captured by supervised loss functions such as the ESR, they can be better dealt with by adversarial losses such as MSD. Compared to MSD only, MSD+MPD can perform even better for such high-gain tones. We speculate that adding MPD helps, for MPD operates directly on equally-spaced sample points of the audio waveform in its original temporal resolution (e.g., 44.1 kHz), while

MSD involves downsampling operations of the waveform. The downsampling operations of MSD may have limited its strength in assessing high-frequency components.

Informal listening also shows that the supervised baseline can already model the tones in EGDB well, and that for these tones the GAN-based models do not offer obvious advantages. MSD+MPD only performs better on ESR for the Marshall tone. However, we note that the GAN-based models are trained under an unpaired setting, which makes it easier to scale up the training data. Future work can further exploit this advantage by applying data augmentation methods or devising more advanced training framework.

6.2. Artifacts Generated by the Proposed Model

During case analysis, we found that the proposed model is still not perfect and there is room for improvement. In particular, the audio generated by the proposed model may exhibit some artifacts. From Figure 3, MSD+MPD generate a splash of harmonics that does not exist in the target mel-spectrogram. It is unclear why the combination of MSD and MPD discriminators cannot detect such artifacts and accordingly prevent the generator from generating them. However, as both MSD and MPD are discriminators operating directly on the audio waveforms, it might be interesting to incorporate spectral-based discriminators that operate on time-frequency representations to seek possible improvement. For example, in neural audio compression task [9], researchers have shown that splitting the STFT into multi sub-bands allows each sub-discriminator focus on specific frequency bands, providing stronger gradient signals to the generator.

Another key observation is that, while both the supervised baseline model and the proposed MSD+MPD model can result in low ESR values, both of them do not yet model the high-frequency components perfectly, especially for the sustain of notes. This suggests that ESR may not be good enough either as a training objective or an objective evaluation metric for amplifier modeling. Future work can explore other auxiliary losses either for the supervised approach or the GAN-based approach.

7. CONCLUSION

In this paper, we have proposed a new GAN-based model for VA modeling by incorporating the MPD discriminator developed in research on neural vocoders. With experiments on two datasets, we showed that the new model leads to improvement across a range of objective metrics over existing supervised and GAN-based models. Moreover, we demonstrated the benefit of a new scenario where combining clean audio from different datasets enhances GAN training, leading to further performance improvement. Following this light, future work can explore more advanced discriminator architectures to reduce model size, speed up training time, or further reduce artifacts. It would also be interesting to apply the GAN-based approach to datasets with greater diversity in musical content, guitar tone, and recording conditions.

8. ACKNOWLEDGEMENTS

The authors would like to thank the support from the Featured Area Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education of Taiwan (113L900901 /113L900902 /113L900903).

9. REFERENCES

[1] Alec Wright, Eero-Pekka Damskägg, Lauri Juvela, and Vesa Välimäki, "Real-time guitar amplifier emulation with deep learning," *Applied Sciences*, vol. 10, no. 3, pp. 766, 2020.

[2] Eero-Pekka Damskägg, Lauri Juvela, Etienne Thuillier, and Vesa Välimäki, "Deep learning for tube amplifier emulation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 471–475.

[3] Christian J. Steinmetz and Joshua D. Reiss, "Steerable discovery of neural audio effects," *arXiv preprint arXiv:2112.02926*, 2021.

[4] Alec Wright, Eero-Pekka Damskägg, and Vesa Välimäki, "Real-time black-box modelling with recurrent neural networks," in *International Conference on Digital Audio Effects*, 2019.

[5] Lauri Juvela, Eero-Pekka Damskägg, Aleksi Peussa, Jaakko Mäkinen, Thomas Sherson, Stylianos I Mimilakis, Kimmo Rauhanen, and Athanasios Gotsopoulos, "End-to-end amp modeling: from data to controllable guitar amplifier models," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[6] Eero-Pekka Damskägg, Lauri Juvela, Etienne Thuillier, and Vesa Välimäki, "Deep learning for tube amplifier emulation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 471–475.

[7] Johannes Imort, Giorgio Fabbro, Marco A Martínez Ramírez, Stefan Uhlich, Yuichiro Koyama, and Yuki Mitsufuji, "Distortion audio effects: Learning how to recover the clean signal," *arXiv preprint arXiv:2202.01664*, 2022.

[8] Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C. Courville, "MelGAN: Generative adversarial networks for conditional waveform synthesis," *Advances in Neural Information Processing systems*, vol. 32, 2019.

[9] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar, "High-fidelity audio compression with improved rvqgan," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[10] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo, "CycleGAN-VC2: Improved cyclegan-based non-parallel voice conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6820–6824.

[11] Yinghao Aaron Li, Ali Zare, and Nima Mesgarani, "StarGANv2-VC: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion," *arXiv preprint arXiv:2107.10394*, 2021.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[13] 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.

[14] Alec Wright, Vesa Välimäki, and Lauri Juvela, "Adversarial guitar amplifier modelling with unpaired data," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[15] Yu-Hua Chen, Wen-Yi Hsiao, Tsu-Kuang Hsieh, Jyh-Shing Roger Jang, and Yi-Hsuan Yang, "Towards automatic transcription of polyphonic electric guitar music: A new dataset and a multi-loss transformer model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 786–790.

[16] Hegel Pedroza, Gerardo Meza, and Iran R Roman, "EGFxset: Electric guitar tones processed through real effects of distortion, modulation, delay and reverb," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[17] Alec Wright and Vesa Välimäki, "Grey-box modelling of dynamic range compression," in *International Conference on Digital Audio Effects*, 2022, pp. 304–311.

[18] Felix Eichas and Udo Zölzer, "Gray-box modeling of guitar amplifiers," *Journal of the Audio Engineering Society*, 2018.

[19] John Covert and David L. Livingston, "A vacuum-tube guitar amplifier model using a recurrent neural network," in *Proceedings of IEEE Southeastcon*, 2013.

[20] Zhichen Zhang, Edward Olbrych, Joseph Bruchalski, Thomas J. McCormick, and David L. Livingston, "A vacuum-tube guitar amplifier model using long/short-term memory networks," *SoutheastCon*, 2018.

[21] Felix Eichas and Udo Zölzer, "Virtual analog modeling of guitar amplifiers with Wiener-Hammerstein models," in *Proceedings of Annual Convention on Acoustics*, 2018.

[22] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[23] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[24] Qingyang Xi, Rachel M Bittner, Johan Pauwels, Xuzhou Ye, and Juan Pablo Bello, "GuitarSet: A dataset for guitar transcription," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 453–460.

[25] Marco Comunità, Dan Stowell, and Joshua D Reiss, "Guitar effects recognition and parameter estimation with convolutional neural networks," *arXiv preprint arXiv:2012.03216*, 2020.

[26] Christian Kehling, Jakob Abeßer, Christian Dittmar, and Gerald Schuller, "Automatic tablature transcription of electric guitar recordings by estimation of score-and instrument-related parameters.," in *DAFx*, 2014, pp. 219–226.

[27] Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim, "Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation," *arXiv preprint arXiv:2106.07889*, 2021.

[28] Jinhyeok Yang, Junmo Lee, Youngik Kim, Hoonyoung Cho, and Injung Kim, "Vocgan: A high-fidelity real-time vocoder with a hierarchically-nested adversarial network," *arXiv preprint arXiv:2007.15256*, 2020.

[29] Jae Hyun Lim and Jong Chul Ye, "Geometric GAN," *arXiv preprint arXiv:1705.02894*, 2017.

[30] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International conference on machine learning*, 2016, pp. 1558–1566.

[31] Christian J. Steinmetz and Joshua D. Reiss, "pyloudnorm: A simple yet flexible loudness meter in python," in *150th AES Convention*, 2021.

[32] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi, "Fréchet audio distance: A metric for evaluating music enhancement algorithms," *arXiv preprint arXiv:1812.08466*, 2018.

[33] Yen-Tung Yeh, Bo-Yu Chen, and Yi-Hsuan Yang, "Exploiting pre-trained feature networks for generative adversarial networks in audio-domain loop generation," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2022.

[34] Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al., "MusicLM: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[35] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez, "Simple and controllable music generation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[36] Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou, "Adapting Fréchet audio distance for generative music evaluation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 1331–1335.

[37] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[38] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, "Least squares generative adversarial networks," in *IEEE International Conference on Computer Vision*, 2017, pp. 2794–2802.

[39] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203.