DISTORTION RECOVERY: A TWO-STAGE METHOD FOR GUITAR EFFECT REMOVAL

*Ying-Shuo Lee**
Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan
r10921a16@ntu.edu.tw

*Yueh-Po Peng**
Institute of Information Science
Academia Sinica
Taipei, Taiwan
yuehpo@iis.sinica.edu.tw

Jui-Te Wu
Positive Grid
Henderson, USA
ray.wu@positivegrid.com

Ming Cheng
Institute of Information Science
Academia Sinica
Taipei, Taiwan
hugowski@iis.sinica.edu.tw

Li Su
Institute of Information Science
Academia Sinica
Taipei, Taiwan
lisu@iis.sinica.edu.tw

Yi-Hsuan Yang
Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan
yhyangtw@ntu.edu.tw

ABSTRACT

Removing audio effects from electric guitar recordings makes it easier for post-production and sound editing. An audio distortion recovery model not only improves the clarity of the guitar sounds but also opens up new opportunities for creative adjustments in mixing and mastering. While progress have been made in creating such models, previous efforts have largely focused on synthetic distortions that may be too simplistic to accurately capture the complexities seen in real-world recordings.

In this paper, we tackle the task by using a dataset of guitar recordings rendered with commercial-grade audio effect VST plugins. Moreover, we introduce a novel two-stage methodology for audio distortion recovery. The idea is to firstly process the audio signal in the Mel-spectrogram domain in the first stage, and then use a neural vocoder to generate the pristine original guitar sound from the processed Mel-spectrogram in the second stage. We report a set of experiments demonstrating the effectiveness of our approach over existing methods, through both subjective and objective evaluation metrics.

1. INTRODUCTION

Electric guitar effects such as distortion usually act as a decisive factor across various musical genres affecting the emotion, color, and the aesthetic taste of music. For many music information retrieval (MIR) tasks, however, such guitar effects add another layer of complexity and can degrade the performance of MIR models. For example, for automatic music transcription, Chen *et al*. [1] found that guitar signals with different pedal effects negatively impact the accuracy of transcription. As such, *distortion recovery*, the task of automatically removing effects from recorded tracks *post hoc*, may provide a solution improving the performance of MIR models, including transcription, source separation and automatic mixing systems [2, 3, 4]. For sound engineers, distortion recovery also make it easier for sound editing.

In prior research, distortion recovery is usually treated as a special case of source separation or source enhancement [5, 6]. Specifically, it is assumed that the distorted signal can be represented as the sum of the clean signal and the "effect signal" (regarded as noise or another source). Signal processing and machine learning methods can then be developed to extract the clean signal from mixed or noisy ones [7, 8]. Although these methods have been shown to be promising, we note that previous research mostly consider only synthetic distortions that may be too simplistic to reflect the complexities seen in real-world recordings. The complex and dynamic features of various effect pedals and Virtual Studio Technology (VST) plugins, combined with the diverse playing styles and recording environments, can all pose challenges for distortion recovery. However, how such nuances in real-world recordings impact the performance of distortion recovery have not been studied thus far, to our best knowledge.

In this paper, we present two contributions to the task of distortion recovery. Firstly, we propose a new technical approach that is inspired by recent advance in voice conversion and synthesis. Specifically, our approach contains two stages. In the first stage, we utilize a "Mel denoiser" which transforms the Mel-spectrogram of the distorted audio signal into that of the non-distorted, dry signal. In the second stage, we employ a neural vocoder to obtain the waveform of the dry signal. Experiments show that, compared to the prevalent single-stage approach, the proposed approach can better reinstate the intricate details inherent in the original guitar recording into the purified waveform. This preservation of the expressiveness and dynamic range of the original signal sets our method apart from the prior arts, demonstrating superior performance in auditory fidelity and processing efficiency.

Secondly, we build and test the implemented models on two distinct datasets: one derived from software simulation using the Pedalboard [9] as done in previous work, and the other from the "BIAS FX2 ToneCloud presets"[1] using commercial-grade VST plugins released by a leading guitar amp and effect modeling company called Positive Grid.[2] This allows for a performance comparison of models in controlled versus real-world environments, offering new insights into the tested models.

*These authors contributed equally to this work

[1] https://www.positivegrid.com/products/bias-fx-2

[2] https://www.positivegrid.com/

Audio samples can be found at our demo page.[3] Moreover, as we use in-house data in our experiments (see Section 4.3), for reproducibility we will create a dataset that can be publicly released (using the dry signals from the EGDB dataset [1]) and report evaluation result on that dataset on the demo page as well.

2. RELATED WORKS

The quality and clarity of audio signals usually play a crucial role in MIR applications such as music transcription and chord recognition. Chen *et al*. [1] assessed the performance of guitar transcription using various settings, including dry Direct Input (DI) signal, and wet signals rendered with amps and real-world recordings sourced from YouTube, observing performance degradation on wet signals compared to the case of dry signals. Pauwels *et al*. [10] also showed that chord recognition datasets often feature clean and well-defined chords, which may not be practical in real-life situations, particularly in guitar signals where distortion effects are commonly used, calling for the need of effect removal.

Distortion recovery is a novel and challenging task lying between the realms of signal processing and machine learning. Deep neural networks (DNNs) have been adopted for distortion recovery lately. Imort *et al*. [5] explored the elimination of distortion and clipping from guitar tracks using various DNN architectures, finding that a model originally developed for source separation [7] works the best. Their work signifies a pivotal shift towards employing deep learning for audio effect manipulation, suggesting the potential of DNNs in distinguishing and isolating the nuanced characteristics of distorted audio signals.

Expanding the scope to encompass general-purpose audio effect removal, Rice *et al*. [6] investigated a scenario with five specific audio effects: distortion, dynamic range compression, reverberation, chorus, and feedback delay. They devised a process named "RemFX," which first detects whether a type of audio effect has been applied, and then removes each effect one at a time. They also showed that source separation-based model such as Demucs V3 [7] and speech enhancement-based model such as DCUNet [8] work well for the removal of audio effects. The goal of the present work is related to but different from theirs—we intend to build a model that removes combinations of multiple correlated distortion effects that are hard to be tackled individually. However, as Demucs V3 and DCUNet have been shown promising in their setting, we adopt these two models as baselines in our experiments.

The neural vocoder, which employs neural networks to reconstruct waveforms from Mel-spectrograms, is widely used in audio processing. Modern neural vocoders, including MelGAN [11], HiFiGAN [12], and iSTFTNet [13], leverage generative adversarial networks (GANs) to achieve high-fidelity results that significantly surpass those of the traditional Griffin-Lim algorithm. In this study, we tackle the task of removing guitar distortion through a two-stage approach: initially processing the Mel-spectrogram, followed by employing a neural vocoder to convert it back into waveform. This method diverges from previous research, which predominantly performs distortion recovery in a single stage directly in the waveform domain (e.g., [7, 8]).

As for data, acquiring data from physical amplifiers presents significant challenges. Juvela *et al*. [14] mechanized the process by attaching electric motors to each pertinent control of a physical amplifier, successfully gathering 4.5 hours of paired signals. However, the utilization of rudimentary algorithms [5] and software applications such as Pedalboard [6, 9] might yield data that are not realistic enough.

[3] https://y10ab1.github.io/guitar_effect_removal/

3. METHOD

3.1. Distortion Recovery Process

The state-of-the-art techniques of audio distortion recovery [5, 6], particularly concerning distortion effects, posit that the mixed signal, y, is represented as a linear blend of *wet* signal $f(x)$ and the *dry* signal x, where the nonlinear distortion function $f(\cdot)$ is applied to x:

$$y = \alpha f(x) + (1 - \alpha)x\,, \tag{1}$$

where $\alpha \in [0, 1]$ represents the influence of the distorted signal. This assumption basically stems from source separation and audio enhancement models [7, 8, 15, 16].

Unlike the prior approach, we instead assume that the distortion effect fundamentally alters the characteristics of the dry component such that it may not be identifiable within the processed output. Distortion typically generates highly nonlinear interactions, not merely attenuating but transforming the dry signal. More specifically, the mixed signal aforementioned can be approximated as a wet signal, y, expressed as follows:

$$y = f^*(x)\,. \tag{2}$$

This realization prompts a shift from traditional linear models to a more sophisticated function that encapsulates the intricacies of this transformation. Inspired by voice conversion and synthesis [17, 18, 19], we articulate Equation 2 as a two-stage model. The first stage focuses on recovering an approximation of the clean signal from the distorted wet signal, acknowledging that it might be devoid of certain fine details:

$$\hat{x}_{\text{approx}} = h(y)\,, \tag{3}$$

where y is the distorted wet signal and $h(\cdot)$ symbolizes the initial recovery function, striving to approximate the clean signal $\hat{x}_{\text{approx}}$. Noting that $\hat{x}_{\text{approx}}$ may lack the subtleties and nuances inherent to the original guitar signal, the second stage is designed to reinstate these details:

$$\hat{x} = g(\hat{x}_{\text{approx}})\,. \tag{4}$$

Here, $g(\cdot)$ is a refinement function tasked with restoring the finer characteristics and nuances to the estimated clean signal $\hat{x}_{\text{approx}}$, yielding the final restored signal $\hat{x}$.

In the following subsections, we introduce the functions $h(\cdot)$ and $g(\cdot)$, referred to as the "Mel Denoiser" and "Neural Vocoder," respectively. Together, they model a two-step restoration process. Initially, $h(\cdot)$ approximates the clean signal from a heavily distorted (or wet) output. Following this, $g(\cdot)$ refines the approximation, polishing it to achieve a high-fidelity restoration of the original dry guitar signal.

3.2. Mel Denoiser: The First Stage of the Proposed Model

To initiate the denoising process, the wet waveform is first transformed into a Mel spectrogram. In our approach, each Mel spectrogram frame is treated as an embedding, effectively converting the Mel spectrogram into a sequence of embeddings. This conversion is ideal for Transformer-based architectures, which excel

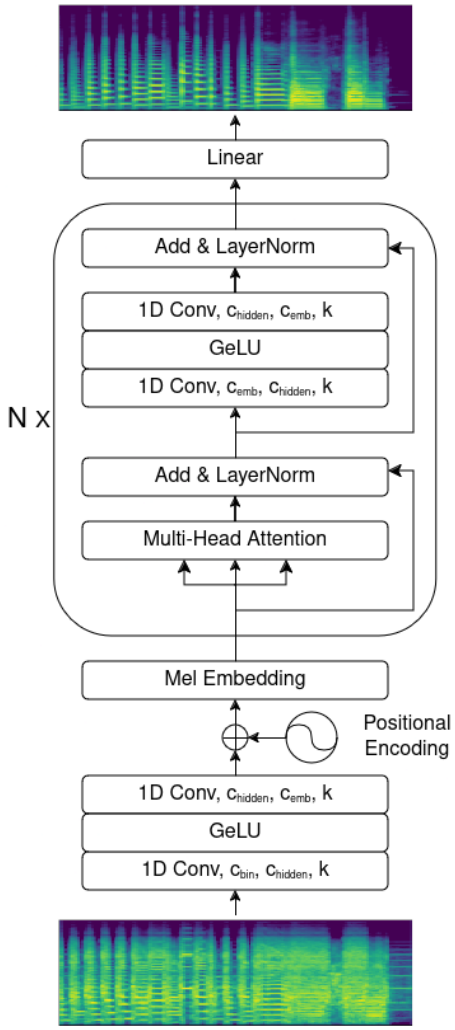

Figure 1: The architecture of the proposed Mel Denoiser. N represents the number of layers, while c_{bin}, c_{hidden}, and c_{emb} indicate the channel counts. The kernel size of the 1D convolution is denoted by k. Here, c_{bin} matches the Mel- spectrogram bin count, and c_{emb} corresponds to the embedding size. We configured c_{hidden} to be four times larger than c_{emb}.

at processing sequences. However, traditional Transformers process the full-length sequence of hidden representations across all layers, resulting in high computational costs. Noting that adjacent frames are most significant for denoising tasks, and drawing inspiration from previous advancements in the text-to-speech (TTS) arena [18], we adopt a pure Transformer encoder with modifications as presented in Figure 1. Specifically, we replace the conventional feed-forward linear layer with two 1D convolution layers, incorporating a GELU activation function. This approach enhances the efficiency and effectiveness of the denoising process. Ultimately, the model is trained to generate a dry Mel spectrogram by removing the unwanted audio effects from the initial input.

3.3. Neural Vocoder: The Second Stage of the Proposed Model

Here we aim to convert the dry signal Mel spectrogram produced by the first stage into a dry waveform. In doing so, we employ the widely-used neural vocoder called HiFi-GAN [20], which leverages generative adversarial networks for waveform generation. The generator within HiFi-GAN takes the Mel spectrogram as input and employs transposed convolution layers to progressively upsample the signal until the length of the output matches that of the target waveform. HiFi-GAN features two discriminators: the multi-period discriminator and the multi-scale discriminator. The former is designed to capture various periodic components of the raw waveforms, while the latter focuses on identifying patterns across different lengths of the raw waveforms, ensuring a rich and accurate audio reproduction.

4. EXPERIMENTS

4.1. Experimental setup

The audio signals are sampled at 44.1 kHz. The Mel-spectrogram is configured with 128 bins, with an window size of 2,048, and a hop length of 512. The Mel Denoiser comprises 12 blocks, each including a self-attention layer with an embedding size of 384, and employs 1D convolutional kernels with size of [9, 1]. For optimization, we use AdamW with a learning rate of 1e–5, a learning rate decay of 0.999999 at each step, and a batch size of 64. The HiFi-GAN implementation is sourced from the vtuber-plan project.[4] The HiFi-GAN is trained with a modified LS-GAN loss, L1 Mel-spectrogram loss, and feature matching loss identical to the original paper, while the Mel Denoiser is only trained with L1 Mel-Spectrogram Loss. Our training regimen begins with separate training of the Mel Denoiser and the Neural Vocoder, conducting 1.5 million steps for the Mel Denoiser and 1 million steps for HiFi-GAN, followed by a "fine-tuning" phase for HiFi-GAN using the output from the Mel Denoiser for an additional 0.5 million steps. All experiments were conducted on a single RTX 4090 GPU.

4.2. Baseline Models

To benchmark the effectiveness of our proposed model in the domain of audio distortion recovery, we draw comparisons with three notable models renowned for their contributions to audio processing tasks: Demucs V3 [7], DCUnet [8], and HiFi-GAN Denoiser [20]. These models are selected for their relevance and demonstrated success in tasks closely aligned with our objectives, such as source separation, denoising, and audio enhancement.

Demucs V3, a.k.a., Hybrid Demucs, is an extension of the U-Net model designed for musical source separation [7]. It combines convolutional layers with LSTM units to capture audio signal dependencies across different scales. It has been adopted in RemFX [6] for distortion removal.

DCUnet is another variant of the U-Net architecture that is designed to work with complex spectrograms by employing complex convolutions. It excels in tasks requiring detailed spectral manipulation, such as speech enhancement and audio denoising, due to its capacity to preserve intricate phase and magnitude information. It has also been adopted in RemFX [6].

HiFi-GAN Denoiser: Distinct from the previously mentioned HiFi-GAN neural vocoder [12], the HiFi-GAN denoiser operates

[4] https://github.com/vtuber-plan/hifi-gan

on a waveform-to-waveform basis, targeting the elimination of a wide array of noises, reverberations, and equalization distortions present in recordings. This denoiser utilizes a feed-forward WaveNet architecture, complemented by discriminators that operate on both the waveform and Mel- spectrogram scales. Additionally, it extracts deep features from the discriminators to enhance the perceptual quality of the audio output. This approach ensures the denoised audio maintains a high level of clarity and fidelity, making the HiFi-GAN Denoiser a strong competitor for our model, especially in terms of maintaining audio quality while removing distortion effects.

Ours-Base: In addition to the set up mentioned in Section 4.1, we trained a variant of our model with fewer trainable parameters for comparison. This base model adopts all the configurations, except it scales down the number of layers from 12 to 8 and reduces the embedding size from 384 to 256.

All the implemented models were trained from scratch for 1.5 million steps using one of the datasets described in Section 4.3, with early stopping to halt training if no advancement in performance was observed. The AdamW optimizer, with a learning rate of 1e–4, facilitated the optimization process. Batch sizes were selected to optimize memory usage and computational efficiency, given the diverse memory requirements across models. The primary training objective for all models was the minimization of the L1 loss on the waveform. Notably, for the HiFi-GAN Denoiser [20], the training also involved L2 loss on log spectrograms and additional adversarial and deep feature matching losses. These losses are particularly effective at capturing and improving perceptual aspects of audio quality, aiming for a denoised output that resembles the natural properties of clean speech signals.

4.3. Datasets

We consider two datasets in our experiments.

VST-derived Data: To have a broad and varied dataset of paired signals, we utilize an in-house dataset containing 80 hours of electric guitar dry and wet signal pairs, provided by Positive Grid. Each clip lasts 4 seconds and is sampled at 44.1 kHz. This dataset complies with Positive Grid's privacy policy and GDPR guidelines, ensuring the protection of personal data and user privacy. Specifically, the dry signals are contributed by 14 professional guitarists under consent agreements, while the wet signals are produced using the BIAS FX2 VST plugins of Positive Grid, with presets randomly selected from its "ToneCloud" library, which includes over 90,000 options. In general, VST plugins often consist of multiple stages of signal processing, including preamp modeling, tone shaping, cabinet simulation, effects processing, and post-processing. Each stage adds complexity to the signal chain and contributes to the overall sound. To focus on distortion-related effects, we consider noise gate, EQ, compressor, overdrive, distortion, and amplifier, while other effects such as delay, modulation, reverb, and pitch shifting are excluded. Notably, while our dataset more accurately reflects typical guitarist recording conditions, discrepancies remain when compared to various real-world sources, such as YouTube recordings captured in different environments.

Synthetic Distortion Data: To enable comparison with prior research, we also created synthetic distortion effects using the Pedalboard [9] on the same dry signals of the previous VST-derived data. Distortion and clipping effects were randomly applied to the dry signals. The gain levels γ for the distortion effect were uniformly chosen from $\gamma \in [20, 50]$ dB, and the clipping thresholds τ were uniformly selected from $\tau \in [-50, -20]$ dB.

4.4. Objective Evaluation Metrics

We consider the following metrics to quantitatively measure the performance of the implemented models.

Fréchet Audio Distance (FAD) [21]: Inspired by the Fréchet Inception Distance (FID) used in computer vision, FAD measures the similarity between distributions of real and synthesized audio features as extracted by a deep learning model. A lower FAD score suggests a higher resemblance to the target audio distribution, indicative of superior audio quality. We report FAD score calculated by pretrained CLAP model [22].

Error-to-Signal Ratio (ESR) [23] is a metric that quantifies the proportion of the error signal relative to the desired signal. It provides an insightful measure of the distortion or unwanted components present in the output audio. Lower ESR values indicate higher fidelity in signal reconstruction.

Scale-Invariant Signal-to-Distortion Ratio (SI-SDR) serves as a robust alternative to the traditional Signal-to-Distortion Ratio (SDR) [24] by providing a scale-invariant measure of signal quality. It is particularly adept at evaluating the degree of distortion removal and the fidelity of the signal reconstruction. Higher SI-SDR values correlate with less audible distortion, suggesting a superior perceptual quality of the recovered audio signal.

Multiresolution STFT (MR-STFT) [25, 26]: In contrast to the conventional STFT where a singular trade-off between time and frequency resolution is inevitable, MR-STFT employs multiple STFT analyses with varied window sizes and resolutions. Doing so enables MR-STFT to capture both fine-grained temporal details and broad frequency characteristics within the same audio signal. Therefore, MR-STFT offers insight into the intricate time-frequency attributes of audio signals.

Number of Parameters: This metric reflects the total number of trainable parameters within the model. A model with fewer parameters is generally more efficient, with a reduced memory footprint and faster inference capabilities.

4.5. Subjective Evaluation Metrics

Subjective evaluations provide a critical measure of an audio processing model's performance, offering insights that objective metrics cannot capture. In this assessment, Mean Opinion Scores (MOS) are employed, where 26 professional guitarists and music producers critically evaluate the perceptual quality of audio and the proficiency of models in restoring distortion-affected signals. To facilitate this, 10 distinct sets of audio samples, not included in the training set, were curated for auditory evaluation. Each set includes five audio files: the unprocessed original *Dry* signal and the outputs processed by Demucs V3, DCUNet, HifiGAN-denoiser, and our model. The participants appraise each audio file, focusing on the quality and the effectiveness of distortion mitigation, on a scale from 1 to 5 (the higher the better).

Audio Quality (AQ): The AQ metric encompasses the overall sound quality after distortion recovery. It reflects the listeners' perception of clarity, fidelity, and the absence of unwanted artifacts. A higher MOS in AQ signifies that listeners perceived the audio as high quality, suggesting effective recovery of the original signal.

Dryness Level (DL): The DL metric evaluates the extent to which the effects, particularly distortion, have been removed from the guitar signal. A higher MOS in DL indicates a signal that

	Params.	FAD ↓	ESR ↓	SISDR ↑	MR-STFT ↓	AQ ↑	DL ↑
Demucs V3 [7]	83.5M	0.383	**0.869**	2.984	2.236	$1.66_{\pm 0.83}$	$1.86_{\pm 0.96}$
DCUnet [8]	7.7M	0.249	0.968	4.085	1.821	$2.10_{\pm 1.11}$	$1.99_{\pm 1.05}$
HifiGAN denoiser [20]	**1.3M**	0.224	1.216	6.212	2.271	$2.67_{\pm 1.12}$	$2.77_{\pm 1.26}$
Ours-Base	45.9M	0.083	2.808	27.608	1.568	—	—
Ours-Large	101.7M	**0.080**	2.290	**28.650**	**1.419**	$\mathbf{3.54}_{\pm 1.10}$	$\mathbf{3.86}_{\pm 1.08}$
Ground Truth	—	—	—	—	—	$4.25_{\pm 0.89}$	$4.29_{\pm 0.95}$

Table 1: The number of parameters and performance of various models trained on VST-derived data. Ours-Large leads to the best result across various objective (middle) and subjective (rightmost) metrics, while having a similar number of parameters as Demucs V3. Ground Truth serves as the high anchor for the two subjective evaluation metrics AQ (audio quality) and DL (dryness level). The arrows ↑ and ↓ indicate the higher or the lower the better; best result in each column is highlighted in bold.

listeners perceive as closely resembling the original, unaffected dry sound, implying that the corresponding model successfully removes the intended effects, restoring the natural state of the signal as heard by the listeners.

5. RESULTS

5.1. Audio Quality and Model Efficiency

To conduct a comprehensive comparison of all models in a realistic scenario, we trained all models with the VST-derived data. Table 1 presents a comparative evaluation of audio quality metrics. We see that the proposed model (i.e., Ours-Large) achieves the lowest FAD score, indicating alignment with the true distribution of the VST-derived data. Moreover, it secures the highest SI-SDR value, reflecting exceptional ability in signal reconstruction fidelity. However, possibly because our model is less sensitive to phase variations, it does not score well on ESR. We note that the ESR metric may not always align with subjective metrics that reflect human perception of music quality. Further research could be beneficial to evaluate the alignment of ESR metrics with human perceptual quality. Demucs V3, despite achieving the lowest ESR score, showing fewer errors in the output signal, falls short compared to our model in other aspects. DCUnet, while computationally efficient, lags behind in performance. The fewer parameters of HiFi-GAN Denoiser hint at a trade-off in audio quality when compared to our more sophisticated model. Our-Base exhibit strong performance with few parameters, with Ours-Large showing exceptional proficiency across all evaluated metrics.

Additionally, we present the Mel-spectrogram of several examples in Figure 2. Our model closely matches the target while baseline models fail to retain high-frequency signals and do not effectively eliminate noise.

5.2. Training with the VST-derived Data vs. with the Synthetic Distortion Data

The analysis of distortion recovery unfolds in two phases: initially with the synthetic data and later with the VST-derived data. These datasets are instrumental in assessing the versatility of model and their capacity to adapt to varying acoustic environments.

Upon training the models with the synthetic data and evaluating them on the VST-derived data, we find that none of the models reach desirable outcomes in terms of the FAD. However, they do achieve relatively high SI-SDR scores. This juxtaposition suggests that with the synthetic data, while fostering high SI-SDR performance, falls short in preparing models for the nuanced complexities encountered in the VST-derived data, as reflected by the elevated FAD values. The FAD metric reveals a significant gap between the model outputs and the practical data, highlighting a potential limitation of the synthetic training data in replicating the full spectrum of the VST-derived audio nuances.

	FAD ↓	ESR ↓	SI-SDR ↑	MR-STFT ↓
Demucs V3 (Synthetic)	**0.375**	2.436	16.792	3.589
DCUnet (Synthetic)	0.392	**1.002**	16.539	2.856
Ours (Synthetic)	0.455	1.827	**31.790**	**2.170**
Demucs V3 (VST)	0.383	**0.869**	2.984	2.236
DCUnet (VST)	0.249	0.968	4.085	1.821
Ours (VST)	**0.080**	2.290	**28.650**	**1.419**

Table 2: Performance comparison of models trained with VST-derived data (VST) and synthetic distortion data (Synthetic). The table highlights our model's superior performance in terms of FAD, SI-SDR, and MR-STFT metrics when trained exclusively on the VST-derived data, in contrast to Demucs V3 and DCUnet, which exhibit lower performance across both types of datasets. This underlines our model's robustness and its enhanced capability to approximate the target dry signal accurately in complicated real-world VST-derived acoustic settings.

Furthermore, when the models are trained with the VST-derived data, our model exhibits a marked improvement over the other models across most metrics. Notably, our model achieves a dramatically lower FAD score (0.080 compared to 0.249 and 0.383 by DCUnet and Demucs V3, respectively) and a significantly higher SI-SDR value (28.650 compared to 4.085 and 2.984 by DCUnet and Demucs V3, respectively) when evaluated on VST data. These results suggest that our model is particularly adept at handling the diverse and complex nature of the VST-derived audio signals, providing a more accurate and reliable removal of effects.

The superior performance of our model on the VST-derived data underlines the importance of training with data that closely mimics the target environment. It also supports the idea that models trained on more representative audio signals are more likely to generalize well to real-world scenarios,[5] confirming the efficacy of our model in managing the unpredictable variations present in

[5] Additional experiments conducted on the EGDB dataset [1] are available on the demo page. The dataset and results can be accessed at `https://y10ab1.github.io/guitar_effect_removal/`.

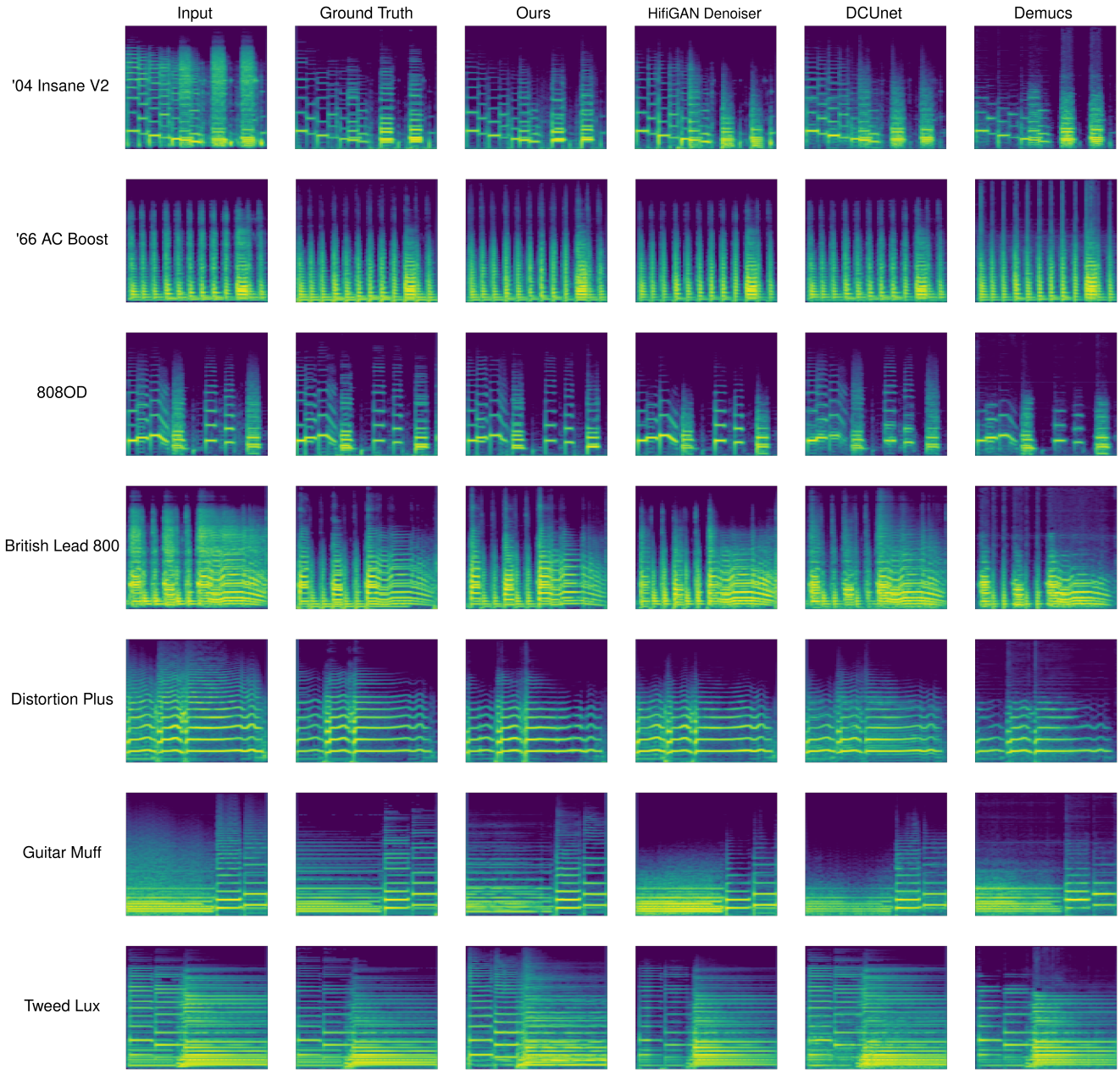

Figure 2: The Mel-spectrograms of the input wet signal, target dry signal, along with the output of the proposed model, the HiFiGAN Denoiser [20], DCUnet [8], and Demucs V3 [7], across a total of seven different VST plugin effects. Our model demonstrates a closer resemblance to the target signal, showcasing superior distortion reduction capabilities and better preservation of overtone characteristics.

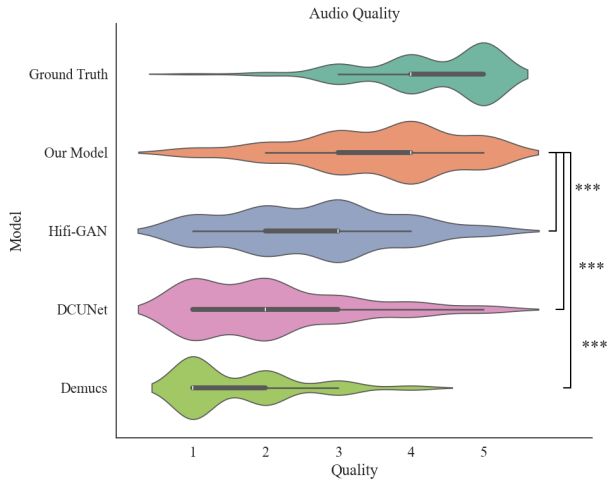

Figure 3: Mean Opinion Scores for Audio Quality (AQ). The distribution indicates that our model primarily achieved ratings around 4 points, signifying a high level of signal quality post-distortion recovery. (*** $= p < .001$ in statistical test).

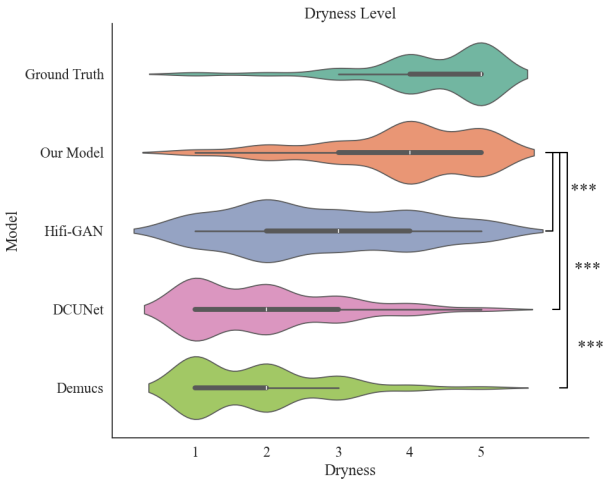

Figure 4: Mean Opinion Scores for Dryness Level (DL). The concentration of ratings around 4 points for our model suggests the dryness of recovered signal is favorably compared to the ground truth, demonstrating effective distortion removal. (*** $= p < .001$.)

	FAD ↓	ESR ↓	SI-SDR ↑	MR-STFT ↓
Ours-Base	0.128	**1.842**	27.532	1.430
+ finetune	0.083	2.808	27.608	1.568
Ours-Large	0.129	1.976	27.802	**1.358**
+ finetune	**0.080**	2.293	**28.651**	1.419

Table 3: Ablation study comparing different model sizes and the effect of vocoder fine-tuning. See Section 5.4 for discussions.

realistic audio environments.

5.3. Subjective Quality Evaluations

Audio quality and dryness level play a crucial role in assessing the performance of distortion recovery models, offering insights into the perceived quality from the listener's perspective. To gauge the effectiveness of our model compared to existing baselines, we conducted a user study focusing on these two aspects.

Figure 3 shows a violin plot of audio quality (AQ) ratings for various models. Our model predominantly received ratings around 4 points, indicating listeners highly regard the recovered signal's quality. This consistent high rating sets a benchmark in distortion recovery. Similarly, Figure 4 displays ratings for the dryness level (DL) of audio signals, reflecting how well the recovered signal matches a clean, undistorted ground truth. The violin plot indicates our model frequently scores 4, demonstrating its effectiveness in removing distortion and maintaining the signal's natural features. Post-hoc analyses using the Tukey HSD test following the ANOVA demonstrated significant differences in MOS ratings between the models. The statistical findings are consistent with objective metrics, showing that our model outperforms Hifi-GAN, DCUNet, and Demucs in terms of audio quality and dryness (all with p-value < 0.001).

5.4. Model Architecture Ablation

Finally, we conducted an ablation study to examine the effects of varying model sizes and the impact of the fine-tuning of the vocoder as described in Section 4.1. According to our informal subjective listening, the larger model (Ours-Large) produces outputs more closely resembling the target, particularly in polyphonic compositions compared to the base model. Additionally, fine-tuning of the vocoder helps in reducing artifacts, yielding outputs that are more realistic from a human perspective. However, the larger model does not outperform the base model in objective metrics; in fact, fine-tuning of the vocoder appears to worsen the ESR and MR-STFT scores. We argue that these metrics may not fully capture human perceptions of sound quality.

6. CONCLUSION

In this paper, we have presented a two-stage methodology for removing audio effects from electric guitar tracks, significantly advancing the state-of-the-art for effect recovery. Leveraging a novel approach that combines Mel-spectrogram purification with neural vocoder-based reconstruction, our model outperforms existing ones in producing high-fidelity original sounds from distorted guitar recordings. Moreover, through a comprehensive evaluation employing a broad mix of VST plugins, we have shown that the proposed model performs well not only for simplistic distortion effects tested in prior works, but also for more complicated VST-derived effects that have not been well studied before.

In future work, we plan to extend our approach to more challenging real-world settings, e.g., on guitar recordings sourced from YouTube. It would also be interesting to apply our model to downstream tasks such as guitar transcription and effect modeling.

7. REFERENCES

[1] 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*, 2022.

[2] M. Martinez Ramirez, Daniel Stoller, and David Moffat, "A deep learning approach to intelligent drum mixing with the Wave-U-Net," Audio Engineering Society, 2021.

[3] Christian J Steinmetz, "Deep learning for automatic mixing: challenges and next steps," in *MDX Workshop at ISMIR*, 2021.

[4] Marco A. Martínez-Ramírez, Wei-Hsiang Liao, Giorgio Fabbro, Stefan Uhlich, Chihiro Nagashima, and Yuki Mitsufuji, "Automatic music mixing with deep learning and out-of-domain data," *arXiv preprint arXiv:2208.11428*, 2022.

[5] 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," in *International Society for Music Information Retrieval Conference*, 2022.

[6] Matthew Rice, Christian J Steinmetz, George Fazekas, and Joshua D. Reiss, "General purpose audio effect removal," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2023.

[7] Alexandre Défossez, "Hybrid spectrogram and waveform source separation," in *ISMIR Workshop on Music Source Separation*, 2021.

[8] Hyeong-Seok Choi, Janghyun Kim, Jaesung Huh, Adrian Kim, Jung-Woo Ha, and Kyogu Lee, "Phase-aware speech enhancement with deep complex U-Net," in *International Conference on Learning Representations*, 2019.

[9] Peter Sobot, "Pedalboard," July 2021, [Online] `https://github.com/spotify/pedalboard`.

[10] Johan Pauwels, Ken O'Hanlon, Emilia Gómez, and Mark B. Sandler, "20 years of automatic chord recognition from audio," in *International Society for Music Information Retrieval Conference*, 2019.

[11] 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*, 2019.

[12] 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*, 2020.

[13] Takuhiro Kaneko, Kou Tanaka, Hirokazu Kameoka, and Shogo Seki, "iSTFTNet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time fourier transform," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[14] 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*, 2023.

[15] Fabian-Robert Stöter, Stefan Uhlich, Antoine Liutkus, and Yuki Mitsufuji, "Open-Unmix - A reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, pp. 1667, 2019.

[16] Jingjing Chen, Qirong Mao, and Dong Liu, "Dual-path Transformer network: Direct context-aware modeling for end-to-end monaural speech separation," in *INTERSPEECH*, 2020.

[17] Benjamin van Niekerk, Marc-Andre Carbonneau, Julian Zaidi, Matthew Baas, Hugo Seute, and Herman Kamper, "A comparison of discrete and soft speech units for improved voice conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022.

[18] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "FastSpeech: Fast, robust and controllable text to speech," *Advances in Neural Information Processing Systems*, 2019.

[19] Jui-Te Wu, Jun-You Wang, Jyh-Shing Roger Jang, and Li Su, "A unified model for zero-shot singing voice conversion and synthesis," in *International Society for Music Information Retrieval Conference*, 2022.

[20] Jiaqi Su, Zeyu Jin, and Adam Finkelstein, "HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," *arXiv preprint arXiv:2006.05694*, 2020.

[21] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi, "Fréchet audio distance: A metric for evaluating music enhancement algorithms," 2019.

[22] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[23] Alec Wright and Vesa Välimäki, "Perceptual loss function for neural modeling of audio systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020.

[24] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey, "SDR – half-baked or well done?," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[25] 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*, 2020.

[26] Christian Steinmetz and Joshua D. Reiss, "auraloss: Audio-focused loss functions in PyTorch," [Online] `https://github.com/csteinmetz1/auraloss`.