DMRN+18: DIGITAL MUSIC RESEARCH NETWORK
ONE-DAY WORKSHOP 2023

QUEEN MARY UNIVERSITY OF LONDON
TUE 19 DECEMBER 2023

Electric Guitar Sound Restoration with Diffusion Models

Ronald K. Mo

School of Computer Science, University of Sunderland, United Kingdom
ronald.mo@sunderland.ac.uk

***Abstract*—This work aims to investigate the potential of employing Denoising diffusion probabilistic models, commonly referred to as *diffusion models*, to revert a processed electric guitar recording to its original, unaltered form while retaining all the expressive elements of the performance such as dynamics and articulation. Specifically, a parallel dataset is constructed, containing both the unprocessed and processed versions of the guitar recordings, which is used for training a diffusion model. To preserve the expressiveness, the model is *conditioned* on the processed guitar recording when restoring the raw guitar sound. This research has the potential to enhance the accuracy of various music information retrieval tasks, such as automatic music transcription.**

I. BACKGROUND

Denoising diffusion probabilistic models, also known as diffusion models, have showcased their capacity for generating realistic images [1]. In particular, a diffusion model comprises two processes. The *forward process* entails the repetitive addition of Gaussian noise to the input data x, such as images. Conversely, the *reverse process* is responsible for denoising a vector sampled from $p(z)$ (i.e., the latent representation of x in an iterative manner, ultimately restoring the input data to its original state. A well-trained diffusion model excels in learning the data distribution $p(x)$ within a provided set of data, enabling it to create novel data and surpass the performance of traditional Generative AI (GenAI) models.

Beyond generating images, diffusion models have found applications in various GenAI tasks [3]. To facilitate the conditioning of the generated content, diffusion models often incorporate a *conditioning mechanism*. Conditional diffusion models are designed to learn the conditional distribution of $p(z|y)$ where y is the *conditioning input* such as class labels, text, or audio. Nevertheless, it's worth noting that the application of GenAI models to audio signal processing remains an area that has not been thoroughly explored, to the best of our knowledge.

II. OVERVIEW

This work seeks to explore the potential of utilizing diffusion models to transform a processed electric guitar recording to its raw format (i.e., a clean electric guitar sound), akin to image denoising. To accomplish this, a parallel dataset containing both the clean and processed guitar sounds is constructed. The clean guitar playing is performed and recorded by the author who is a professional studio guitarist. The recorded guitar recording is processed using *Logic Pro*. Due to the preliminary nature of this study, it only considers *distortion* as the processing technique.

A diffusion model is developed and trained using the dataset mentioned above. To reduce the training time, a *latent* diffusion model is used [4]. It first *encodes* the input x (i.e., $\mathcal{E}(x)$) into a lower-dimension representation, carries out both the forward and reverse processes on $\mathcal{E}(x)$, and eventually *decodes* $\mathcal{E}(x)$ (i.e., $\mathcal{D}(\mathcal{E}(x))$) to obtain $\tilde{x}$. To preserve the expressiveness of the guitar playing, the model conditions its output on the *processed guitar recording*. More precisely, the conditional input y is encoded using the Diffusion Magnitude-Autoencoding introduced by Schneider et al. [5] and concatenated with the sampled vector during the reverse process. The generated outputs will be evaluated both objectively and subjectively.

III. CONCLUSION

While the full potential of diffusion models in audio signal processing remains largely unexplored [6], this study introduces a novel approach for the restoration of electric guitar sounds using diffusion models. Given the potential to extend this method to multi-track scenarios, this research has the capacity to enhance the performance of various music information retrieval tasks, including automatic music transcription, music source separation, and beat detection.

REFERENCES

[1] J. Ho et al. "Denoising diffusion probabilistic models," in *Advances in neural information processing systems 33*, 2020, pp. 6840–6851.

[2] M.W.Y. Lam et al. "BDDM: Bilateral Denoising Diffusion Models for Fast and High-Quality Speech Synthesis," in 2022-*10th International Conference on Learning Representations*, 2022.

[3] G. Mittal et al. "Symbolic Music Generation with Diffusion Models", in *Proc. of the 22nd Int. Society for Music Information Retrieval Conf.*, 2021, pp. 468-475

[4] R. Rombach et al. "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684-10695.

[5] F. Schneider et al. "Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion," in *arXiv preprint arXiv:2301.11757,* (2023).

[6] Kong, Zhifeng, et al. "Diffwave: A versatile diffusion model for audio synthesis. " in *arXiv preprint arXiv:2009.09761* (2020).

Ronald K. Mo is with the School of Computer Science. University of Sunderland, United Kingdom (corresponding author e-mail: ronald.mo@ sunderland.ac.uk).