FOLEY SOUND SYNTHESIS IN WAVEFORM DOMAIN WITH DIFFUSION MODEL

Technical Report

Yoonjin Chung[1], *Junwon Lee*[1], *Juhan Nam*[1,2],

[1] Graduate School of AI, KAIST
[2] Graduate School of Culture Technology, KAIST
{yoonjin.chung, james39, juhan.nam}@kaist.ac.kr

ABSTRACT

Foley sound synthesis becomes an important task due to the growing popularity of multi-media content, which is an industrial use-case of general audio synthesis. We propose a diffusion-based model that generates class-conditioned general audio in a classifier-free guidance manner as a participant of DCASE 2023 challenge task 7[1]. Our model follows a UNet-like structure while incorporating LSTM[2] inside the encoder block. We demonstrate the FAD(Frechet Audio Distance) scores of generated results for each 7 sound class respectively.

***Index Terms*—** Diffusion, foley sound synthesis, general audio synthesis, waveform generation, classifier-free guidance

1. INTRODUCTION

The development of deep neural networks has allowed for high-fidelity audio that mimics everyday sounds(e.g. bird chirping, dog barking, or rain) to be produced by promising generative models. This is commonly referred to as general audio synthesis. There has been an increased interest in audio synthesis using deep generative models for personalized sound generation, particularly for multimedia content. In particular, foley sound synthesis is one of the most important production tasks in the posterior phase which aims to generate a sound aligned with the video [1]. The quality of the sound (e.g. appropriate timber, loudness) matters in real use cases because the task is part of the production procedure.

To synthesize proper sounds from various conditions, the studies usually focus on generating audio based on a corresponding informative local conditioner(e.g. mel-spectrogram or aligned linguistic features), or with a single type of source. Example applications of this research include text-to-speech [3] and focusing on specific sound sources like footsteps [4], laughter [5] and drum [6, 7]. A few recent studies propose source-agnostic systems that can synthesize various types of general audio [8, 9, 10]. However, due to the vast diversity and complexity of general audio, most previous works struggle to predict all possibilities of sound sources and appearance.

In another approach, some studies have attempted to synthesize general audio using auto-regressive models [11, 12] and diffusion models [7, 13]. DAG, which is the state-of-the-art in full-band general audio synthesis [13], especially incorporated an autoregressive module inside the diffusion model to tackle high-quality general audio synthesis.

In this work, we exploit the main idea of DAG and propose a diffusion-based generative model in the waveform domain, while leveraging LSTM [2] in the latent space. Our architecture is capable of obtaining various types of sounds with sound category as a condition. We demonstrate our approach enables general audio synthesis on a foley sound dataset without potential quality loss and pretraining. We follow the details given in Foley Sound Synthesis Challenge Track B (task 7) of DCASE 2023 as one of the participants [1].

2. RELATED WORKS

2.1. General audio synthesis with Neural Models

To tackle generative audio generation, two branches of approach have been explored in terms of the model architecture. One is to exploit autoregressive models: SampleRNN based apporach [11], PixelSNAIL with VQ-VAE-2 [12]. The non-autoregressive scheme, especially diffusion models, is the other direction. Score-based diffusion models have been actively explored recently in various sound generation tasks due to their high performance: neural vocoder [14, 15], conditional drum sound generation [7], general audio generation [13], text-to-audio generation [9, 10], etc.

Audio generation schemes can also be divided into two in another direction. The model can either predict time-frequency representation [11, 12, 9, 8, 10] or raw waveform of the sound [14, 15, 7, 13]. In the case of dealing with time-frequency representation such as mel-spectrogram, it requires an auxiliary vocoder module to predict the phase of each frequency component other than frequency magnitude. This often requires pretraining and brings information loss depending on its performance [13]. Generating audio waveform directly, on the other hand, can prevent these potential performance losses.

Audio generation in latent space is a technique proposed to ease training by decreasing the dimension of learning space [9, 10]. The latent space could be rather discrete [12, 9] or continuous [10] as the depth and dimension are also a hyperparameter. However, this method can degrade the generation performance due to information loss while compression [13].

2.2. Diffusion-based models for general audio synthesis

Several previous works attempted to generate a raw waveform of general audio by diffusion-based model. Two score-based models for general sound generation were proposed for the first time: Wavegrad [14] and DiffWave [15] They focus on neural vocoding with mel-spectrogram conditioning by exploiting diffusion probabilistic models to convert mel-spectrograms into raw audio waveforms for speech synthesis. Unfortunately, the two lack the ability to generate sound from scratch which is crucial in general audio synthesis.

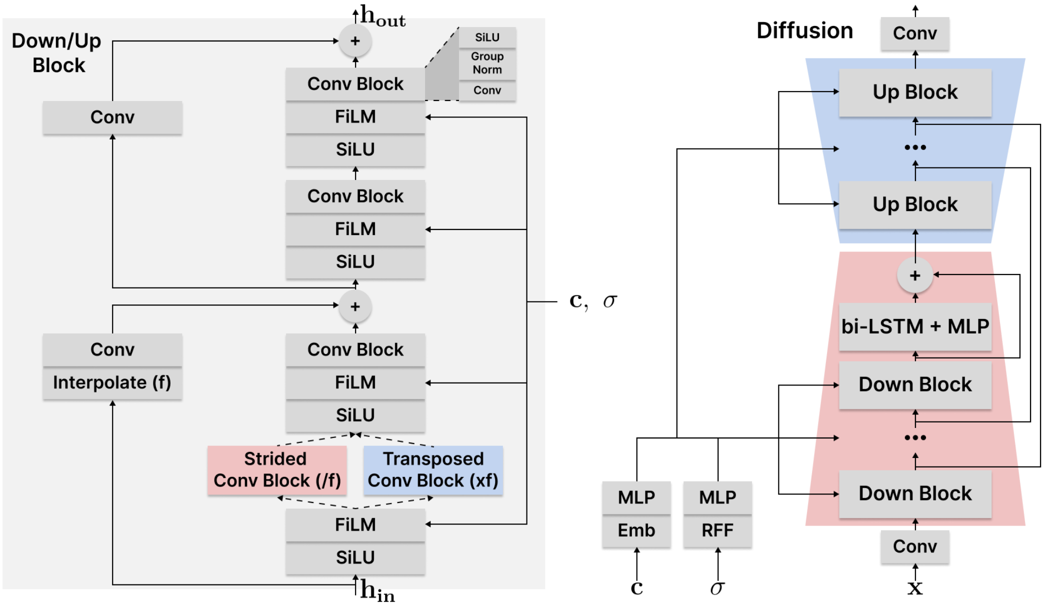

Figure 1: Model architecture.

CRASH [7] is the first approach to generate a waveform of categorical sounds. Their objective is to generate 0.5-second audio waveform of three categories (kick, snare, and cymbals) in a classifier guidance manner. The limitations of this work are the short generation length and the dependency on the auxiliary classifier model.

DAG [13] is the state-of-the-art general audio synthesis system, which outperforms both SampleRNN [11] and an approach with PixelSNAIL and VQ-VAE [12]. The diffusion model is trained on 10 or 15 different sound classes depending on the dataset, with a length of 4 seconds. To implement classifier-free guidance for class-conditioned generation, a learnable class embedding is fed through a FiLM [16] layer with the diffusion time step information. The notable part is that DAG exploited neither the time-frequency representation of audio nor latent space while training.

3. MODEL ARCHITECTURE

Figure 1 shows the model architecture. We follow the idea from DAG [13] to use a UNet-like structure that includes autoregressive module in the encoder block. To predict the noises corresponding to each diffusion step, our proposed conditional UNet model consists of three modules: downsample, middle-sequential, and upsample modules. Downs and up-sampling modules follow the encoder-decoder format, compressing the noised waveform into hidden feature embeddings and vice versa through subsequent layers with resizing factors. Each down/up layer, which refers to the Down/Up Block in Figure 1, contains 4 sequences of activation function, FiLM layer, and convolution block with 2 residual connections. There is one difference between the Down and Up blocks. In the Down block, the first convolution block is built as a stride convolution block with compressive stride factors, while in the Up block, it is built as a transposed convolution block with corresponding factors. Note that every FiLM layers take sigma embedding and class embedding as conditions.

Upon conducting experiments in this basic UNet architecture, we discovered that a lack of global consistency exists in the generated samples. Although the generated samples successfully captured localized features for each class, there was a lack of natural coherence among the sounds. For instance, in the ”Keyboard” class, because the dataset has both typewriter and computer keyboard sounds, these two distinct sounds are blended within a single sample highlighting the presence of certain shortcomings. To address this issue, we added a bidirectional LSTM [2] as a middle-sequence module to achieve a broader receptive field and finally improve the consistency of generated sounds.

For noise scheduling, variance preserving cosine scheduling is used [7], which can be represented as $\sigma(t) = \frac{1}{2}[1 - cos(\pi t)]$ where $\sigma^2(t)I$ stands for the variance of a normal distribution which is the transition kernel of a forward process $p_t(\mathbf{x}(t)|\mathbf{x}(0))$. We also sample t in the interval $[\eta, 1]$ during the training where $\eta = 10^{-4}$.

For conditioning elements, we first embed the logarithm with random Fourier feature embeddings same as [7], followed by a multi-layer perceptron(MLP), to yield the embeddings g in Figure 1. To obtain class condition embedding c, the class label index is converted into an embedding matrix $L \times 512$ where L is the size of class labels and passed to linear projection layers.

4. EXPERIMENTAL DETAILS

4.1. Dataset

We use the provided development data resources only allowed for Track B. The dataset is composed of approximately 5k foley sound samples. Samples have a 22050Hz sample rate in mono for 4 seconds. We used about 95% of the dataset to train and 5% to validate our model.

4.2. Configurations

Through the various experiments with different designs and configurations, we found the best combinations of hyper-

Class ID	Category	FAD	
		Baseline	Ours
0	DogBark	13.411	**8.441**
1	Footstep	8.109	**7.761**
2	GunShot	7.951	**7.892**
3	Keyboard	5.230	**5.167**
4	MovingMotorVehicle	**16.108**	16.358
5	Rain	13.337	**13.173**
6	Sneeze/Cough	**3.770**	4.418
	Total Average	9.702	**9.03**

Table 1: Evaluation result of given sound category classes. FAD stands for Frechet Audio Distance.

parameters which has $[2, 2, 3, 3, 5, 5, 7]$ factors, with channel sizes $[64, 128, 128, 256, 256, 512, 512]$. Therefore, with 22050Hz waveforms, we obtain latent sequences of 512 dimensions at 3.5Hz. The model is trained for 270 iterations on the provided development dataset via gradient descent with Adam, 4 sizes of mini-batch, and a scheduled learning rate from 2×10^{-4} to 5×10^{-5}.

5. RESULTS

For quantitative evaluation, we used the Fréchet audio distance(FAD) [17] score as follows as the challenge's criteria. FAD scores per class are measured from our generated sounds and the given evaluation data on 100 samples for each. is shown in Table 1. Our model outperforms the given baseline [12] in 5 classes and shows similar performances in the rest of the 2 classes.

In terms of qualitative evaluation, we found that the timber of the generated audio is consistent within a single sample, and the quality is better than the baseline model except for a few classes (MovingMotorVehicle and Sneeze/Cough).

6. CONCLUSION

In this work, to address the given challenge of foley sound synthesis, we extended the approach of DAG [13] and proposed a score-based waveform generative model leveraging the temporal features of audio with additional sequential modeling. Through the experiments, we find that foley sounds can be roughly divided into event-driven sounds(e.g. dog bark, sneeze) and environmental sounds(e.g. rain, moving motor vehicle). In future studies, we will focus on the differences between these two types of sounds and study event controllable sound generation.

7. ACKNOWLEDGEMENT

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program(KAIST))

8. REFERENCES

[1] K. Choi, J. Im, L. Heller, B. McFee, K. Imoto, Y. Okamoto, M. Lagrange, and S. Takamichi, "Foley sound synthesis at the dcase 2023 challenge," *In arXiv e-prints: 2304.12521*, 2023.

[2] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *arXiv preprint arXiv:1402.1128*, 2014.

[3] P. Taylor, *Text-to-speech synthesis*. Cambridge university press, 2009.

[4] M. Comunità, H. Phan, and J. D. Reiss, "Neural synthesis of footsteps sound effects with generative adversarial networks," *arXiv preprint arXiv:2110.09605*, 2021.

[5] M. M. Afsar, E. Park, É. Paquette, G. Gidel, K. W. Mathewson, and E. Muller, "Generating diverse realistic laughter for interactive art," *arXiv preprint arXiv:2111.03146*, 2021.

[6] J. Nistal, S. Lattner, and G. Richard, "Drumgan: Synthesis of drum sounds with timbral feature conditioning using generative adversarial networks," *arXiv preprint arXiv:2008.12073*, 2020.

[7] S. Rouard and G. Hadjeres, "Crash: raw audio score-based generative modeling for controllable high-resolution drum sound synthesis," *arXiv preprint arXiv:2106.07431*, 2021.

[8] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "Audiogen: Textually guided audio generation," *arXiv preprint arXiv:2209.15352*, 2022.

[9] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[10] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023.

[11] Q. Kong, Y. Xu, T. Iqbal, Y. Cao, W. Wang, and M. D. Plumbley, "Acoustic scene generation with conditional samplernn," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 925–929.

[12] X. Liu, T. Iqbal, J. Zhao, Q. Huang, M. D. Plumbley, and W. Wang, "Conditional sound generation using neural discrete time-frequency representation learning," in *2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2021, pp. 1–6.

[13] S. Pascual, G. Bhattacharya, C. Yeh, J. Pons, and J. Serrà, "Full-band general audio synthesis with score-based diffusion," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[14] N. Chen, Y. Zhang, H. Zen, R. J. Weiss, M. Norouzi, and W. Chan, "Wavegrad: Estimating gradients for waveform generation," *arXiv preprint arXiv:2009.00713*, 2020.

[15] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "Diffwave: A versatile diffusion model for audio synthesis," *arXiv preprint arXiv:2009.09761*, 2020.

[16] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, “Film: Visual reasoning with a general conditioning layer,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[17] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, “Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms.” in *INTERSPEECH*, 2019, pp. 2350–2354.